Summer 2018

Extended Poisson Models for Count Data With Inflated Frequencies

Monika Arora
Old Dominion University, monikaarora.stats@gmail.com

EXTENDED POISSON MODELS FOR COUNT DATA WITH INFLATED FREQUENCIES

by

Monika Arora
B.Sc. May 2007, University of Rajasthan, India
M.Sc. May 2011, Indian Institute of Technology, Mumbai, India
M.S. December 2015, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTATIONAL & APPLIED MATHEMATICS

OLD DOMINION UNIVERSITY
August 2018

Approved by:

N. Rao Chaganty (Director)

Norou Diawara (Member)

Kayoung Park (Member)

Tina Cunningham (Member)

ABSTRACT

EXTENDED POISSON MODELS FOR COUNT DATA WITH INFLATED FREQUENCIES

Monika Arora
Old Dominion University, 2018
Director: Dr. N. Rao Chaganty

Count data often exhibits inflated counts for zero. There are numerous papers in the literature that show how to fit Poisson regression models that account for the zero inflation. However, in many situations the frequencies of zero and of some other value k tends to be higher than the Poisson model can fit appropriately. Recently, Sheth-Chandra (2011), Lin and Tsai (2012) introduced a mixture model to account for the inflated frequencies of zero and k. In this dissertation, we study basic properties of this mixture model and parameter estimation for grouped and ungrouped data. Using stochastic representation we show how the EM algorithm can be adapted to obtain maximum likelihood estimates of the parameters. We derive the observed information matrix which yields standard errors of the EM estimates using ideas from Louis (1982). We also derive asymptotic distributions to test significance of the inflation points. We use real life examples to illustrate the procedure of fitting our model via EM algorithm.

The second part of this dissertation deals with a generalization of this mixture model where the one parameter Poisson distribution is replaced by a two parameter Conway-Maxwell-Poisson (CMP) distribution, which unlike the Poisson distribution accounts for both over and underdispersion in the count data. The CMP distribution has recently gained popularity, and a CMP model for zero inflated count data was introduced by Sellers and Raim (2016). We discuss properties of the CMP distribution and propose a new mixture distribution, namely zero and k inflated Conway-Maxwell-Poisson (ZkICMP) to address inflated counts with over and underdispersions. We develop regression models based on ZkICMP and discuss parameter estimation using analytical and numerical methods. Finally, we compare goodness of fit of inflated and standard models on simulated and real life data examples.

I dedicate this thesis to my mom Surendra Arora and my boyfriend Vikas Vikram Singh. It is their support, unconditional love and believe in me that I could face the challenges of my graduate school.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere thanks and gratitude to my advisor Professor N. Rao Chaganty for his numerous suggestions and patient guidance throughout the development of this dissertation. His help and encouragement have proven essential both to the completion of this work and to my academic development. For these reasons I am eternally grateful. I am thankful to him for inspiring me to always be enthusiastic to learn, be helpful and kind to the peers. He has been my mentor in my professional and personal development. I am eternally grateful to him for teaching me so much about Statistics and life in general.

I am also immensely grateful to Dr. Tina Cunningham, Dr. Norou Diawara, and Dr. Kayoung Park for serving on my dissertation committee. Their insightful review and expertise have improved my work.

I am also grateful to Dr. Hideaki Kaneko, Chair of the Department of Mathematics and Statistics, and Dr. Raymond Cheng, the Graduate Program Director, for their kind support. I am thankful to the department and Benthic Ecology lab for the financial support. I am grateful to Dr. Daniel Dauer and Mr. Michael Lane for providing a warm and cheerful environment in the lab. I would also like to thank our office managers Ms. Sheila Hegwood and Ms. Miriam Venable. I thank all my fellow graduate friends at the department. Some of them have graduated and some are soon going to graduate. Their company and our discussions on research, graduate life has always brought me immense joy.

I thank my parents, it is them who pioneered my interests in Science and Maths. It is their and my sibling Sonika's love and sacrifice that I could fulfill my aspirations. My boyfriend Vikas has been my pillar during my degree. I am thankful for his support and trust in me. My friends, Nityasri Mandayam (Dodo) and Devan Conroy have always been my family. They have always encouraged me and picked me up whenever I was down. I would also like to acknowledge my friends Amanda Working, Diane Rucci and John Asija who made my stay in Norfolk smooth and comfortable. I can't thank them all enough to be always being there for me. There have been innumerable friends and mentors on the journey so far. I will always be grateful to those who ever inspired and taught me something.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 BACKGROUND

1.1.1 POISSON MODELS

Data that counts the number of occurrences of certain events, or the number of subjects or items that fall in certain categories arise in many scientific investigations, medical and social science research. The most commonly used models to analyze such data are developed using Poisson probability distribution. The Poisson distribution possesses the equidispersion property because the variance which is a measure of dispersion is equal to the mean for this distribution. However, in real life examples most often the data is overdispersed or underdispersed. The common solution for handling overdispersion or underdispersion is to replace the Poisson distribution with a negative binomial or the simpler geometric distribution.

There could be several reasons that lead to overdispersion in the data. A primary cause of overdispersion in the count data is an inflated number of zeros in excess of the number expected under the Poisson distibution. In such cases, an appropriate model is the zero inflated Poisson (ZIP). There are numerous papers in the literature dealing with the ZIP model. The earliest paper on the ZIP model was by Cohen (1960). In a seminal paper, Lambert (1992) introduced and studied the ZIP regression model using Expectation- Maximization (EM) approach. The ZIP model with random effects has been studied by ((Yau and Lee, 2001), (Min and Agresti, 2005)). Ghosh et al. (2006), explored the Bayesian approach for small to moderate sample sizes. The ZIP models using Bayesian approach for spatial data were studied by Agarwal et al. (2002). Further, ZIP models for censored data were studied by Saffari and Adnan (2011) and Yang and Simpson (2012).

In health science research, zero inflated count models have been shown to perform better than traditional count models, for example, see the articles by Umbach (1981), Gupta et al. (1996), and Yau and Lee (2001). ZIP models have been applied across a wide spectrum of academic disciplines including biology (Ridout et al., 1998), ecology (Welsh et al., 1996), psychology ((Atkins and Gallop, 2007), (Loeys et al., 2012)) and education ((Salehi and Roudbari, 2015)). The ZIP models have been studied in economics by Cameron and Trivedi (2013), Greene (1994), Gurmu and Trivedi (1996). In industry, the ZIP models have been applied in manufacturing and transportation. In manufacturing, Lambert (1992) applied ZIP to the number of defects in a manufacturing process with covariates like masking, soldering, etc. See Ghosh et al. (2006) for another application of the ZIP model in manufacturing. Qin et al. (2004) and Lord et al. (2005)) illustrate the use of ZIP models in transportation. A good review and applications of ZIP models is given in Ridout et al. (1998) and Bohning and Seidel (2003). The other ZIP like models are zero inflated negative binomial (ZINB), zero inflated geometric (ZIG), and zero inflated Binomial (ZIB). Numerous authors have investigated these models. For example, Hall (2000) illustrated use of ZIP and ZIB in horticulture.

There are two procedures in SAS that deal with zero inflated models. These procedures are new and in experimental stages. Both, the finite mixture model (FMM) and count regression (COUNTREG) procedures function like the glm procedure and provide estimates, standard errors, and AIC values. However, tests for the mixing proportions are unreliable. The high-dimensional count regression procedure (HPCOUNTREG) in SAS can handle big data. In R, the package 'pscl' includes functions for handling zero inflated discrete distributions with various link options. The inflated count models are also available in the 'VGAM' package.

In addition to zero, some data sets may have an inflated counts of additional value k as a result of multiple effects including the design of the study. Research questionnaire studies are examples with zero and k inflated count data sets typically as a result either in the way the questions were asked or the way the responses were provided. For example, one study investigating the frequency of pap smear tests in women for last six years. The survey had large number of women who never had a pap smear and many who had pap smears on an annual basis. Thus, the survey resulted in a large frequencies of zero and six. The other source for inflation is the

nature of the response. For example, consider the study that counts the number of days per week a subject felt depressed in a sample that consists of depressed and non-depressed subjects. For several non-depressed the count will be zero and for many depressed the count will be 7. Thus the data will likely to have 0 and 7 counts inflated. Lin and Tsai (2012) describe a survey where adults were asked about the number of cigarettes they consume on a given day. The responses tend to be none or a pack. Since a pack consists of 20 cigarettes, the data results in inflated frequencies for 0 and 20. Lin and Tsai (2012) proposed a zero and k inflated Poisson regression model (ZkIP) to analyze such data. In a PhD dissertation, Sheth-Chandra (2011) also introduced two forms of ZkIP models, known as doubly inflated Poisson (DIP) models. In this dissertation, we study the ZkIP form given by Lin and Tsai (2012). It is the same as the second DIP model proposed by Sheth-Chandra (2011).

The ZkIP is a finite mixture model. It has three components. The first is degenerate at zero with probability π_1. The second distribution is degenerate at k with probability π_2 and the third distribution is Poisson with mean λ with probability $\pi_3 = (1 - \pi_1 - \pi_2)$. The mixture leads to heterogeneity in the data which is not captured by the Poisson model. These components can also be interpreted as three groups of the population. A special case of ZkIP model is the zero and one inflated Poisson model (ZOIP). Recently, Zhang et al. (2016) studied the properties and inference on the parameters of the ZOIP distribution without covariates. The inference of ZOIP without covariates was described by Alshkaki (2016). A Bayesian approach for the ZOIP model was examined by Tang et al. (2017). Lin and Tsai (2012), introduced the ZkIP regression model and used the non-linear optimization method to obtain the maximum likelihood (ML) estimates and standard errors. The ZkIP has also been studied by Finkelman et al. (2011) for grouped psychological data. In this dissertation we study the ZkIP model using the Expectation-Maximization (EM) approach. Further we pursue the method outlined by Louis (1982) to obtain the standard errors for the EM parameter estimates.

1.1.2 CONWAY-MAXWELL-POISSON MODELS

As mentioned in Section 1.1.1, the Poisson distribution is the most commonly

used distribution for analyzing count data. The popularly used alternate for overdispersed data is the negative Binomial distribution. However, this distribution does not account for underdispersion in the data. A generalization of the Poisson distribution that can handle both under and overdispersion is the Conway-Maxwell-Poisson (CMP) distribution that was first published by Conway and Maxwell (1962). The CMP is a two parameter (λ, ν) extension of the Poisson distribution. The parameter λ is the rate parameter and ν is the dispersion parameter. The distribution belongs to the exponential family. Further, it turns out not only Poisson but also Bernoulli and geometric distributions are special cases of the CMP distribution. Shmueli et al. (2005) studied extensively the distributional properties of the CMP distribution.

There are numerous papers in the literature that demonstrate the wide applicability of the CMP distribution. Lord et al. (2008) used the CMP distribution to model the number of motor vehicle crashes. Regression models using CMP were used by Kadane et al. (2006) in health care research, and by Shmueli et al. (2005) in marketing, Telang et al. (2004) in eCommerce, Ridout and Besbeas (2004) in biology. Rodrigues et al. (2009) and Balakrishnan and Pal (2015) studied cure rate models using CMP distribution. Bayesian analysis of CMP models was undertaken by Kadane et al. (2006). A review of various extensions of CMP models and their applications can be found in Sellers et al. (2012).

In a recent paper, Sellers and Raim (2016) introduced an extension ZICMP of CMP distribution to study zero inflated count data. It is a mixture of a degenerate distribution at zero with probability π and CMP(λ, ν) distribution with probability $(1 - \pi)$. The special cases of ZICMP are zero inflated Poisson (ZIP), zero inflated bernoulli (ZIB), and zero inflated geometric (ZIG) distributions. The ZICMP model has been used in psychology (Sellers and Raim, 2016), health (Choo-Wosoba et al., 2016), and agriculture (Barriga and Louzada, 2014) research. The statistical software SAS includes CMP and ZICMP in Proc COUNTREG, and Proc HPCOUNTREG.

In this dissertation, we introduce an extension ZkICMP of the ZICMP that accounts for inflated frequencies at zero and k. The ZkICMP is a generalization of ZkIP. It is a mixture of three distributions, first distribution is degenerate at zero with probability π_1, the second is degenerate at count k with probability π_2 and the third is CMP(λ, ν) with probability $\pi_3 = 1 - \pi_1 - \pi_2$. The proportion of zeros which are not from CMP distribution are π_1, while the proportions of k's not belonging to

CMP distribution are π_2. The special cases of ZkICMP are ZkIP, zero and k inflated geometric (ZkIG), zero and k inflated Bernoulli (ZkIB).

1.2 OVERVIEW OF THE DISSERTATION

This dissertation is organized as follows. In Chapter 2, we discuss zero and k inflated Poisson models for the grouped data. We describe in detail two methods for parameter estimation, maximum likelihood (ML) and expectation-maximization (EM). Further, we give formulas to obtain the standard errors for the ML estimates via Fisher information. For the EM estimates we outline the method originally suggested by Louis (1982) to get the standard errors. Further, we show the asymptotic distribution of the likelihood ratio statistic is a mixture of chi-squares, when the mixing parameter falls on the boundary. This result is used to find the significance for testing whether ZkIP could be reduced to ZIP or ZIP could be reduced to a Poisson model. We illustrate our theoretical results using two real life data examples obtained from the National Health Interview Survey (NHIS).

In Chapter 3, we study the zero and k inflated Poisson regression model. We describe the EM method of estimation for estimating the regression parameter and follow the method of Louis (1982) to obtain the standard errors of the estimates. We illustrate our results on the two data sets that we have used in Chapter 2.

In Chapter 4, we construct a zero and k inflated CMP (ZkICMP) regression model. We present properties of the ZkICMP distribution. We also derive formulas for ML estimates of the parameters. Details for obtaining the standard errors via the Fisher information matrix are given in the chapter. We use two data sets from National Health and Nutrition Examination Survey (NHANES) and perform two simulations to demonstrate the importance of ZkICMP model to capture double inflation and over- or underdispersion.

In Chapter 5, we provide a summary and briefly describe possible extensions and future directions of our research.

CHAPTER 2

ZERO AND K INFLATED POISSON MODELS

2.1 INTRODUCTION

In this chapter we study the zero and k inflated Poisson (ZkIP) model for the data that consists of frequencies for counts. We assume the data do not include covariate measurements and are expressed in the form of a frequency distribution. The outline of this chapter is as follows. In Section 2.2, we describe the ZkIP distribution and some of its properties. We present the likelihood function for the ZkIP model in Section 2.3. In Section 2.4, we discuss two methods for parameter estimation, namely, the maximum likelihood estimation and expectation and maximization (EM) method (Dempster et al., 1977). We give details of the Fisher information which provides the asymptotic standard errors for the maximum likelihood estimates in Section 2.5.1. In Section 2.5.2 we describe the method first described by Louis (1982) on how to find the standard errors for the EM estimates for the ZkIP model. In the last Section 2.7, we illustrate the application of our ZkIP model, parameter estimation and calculation of standard errors using two real life examples.

2.2 ZKIP PROBABILITY DISTRIBUTION

The probability mass function of a count variable Y that follows a Poisson distribution is given by

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots \text{ and } \lambda > 0.$$

A model that accounts for the inflated probability at zero is obtained by mixing the Poisson distribution with a point mass π_1 at zero. The probability density function

of this model is given by

$$P(Y=y)=\begin{cases}\pi_1+(1-\pi_1)e^{-\lambda} & \text{when } y=0\\ (1-\pi_1)\dfrac{\lambda^y e^{-\lambda}}{y!} & \text{when } y\geq 1.\end{cases}$$

where $0<\pi_1<1$ and $\lambda>0$. In addition if the probability is also inflated at another count value k, we could consider a Poisson distribution that is mixed with two point masses π_1 and π_2 at 0 and k respectively. In this case the probability mass function of Y is a mixture distribution and it is given by

$$P(Y=y)=\begin{cases}\pi_1+(1-\pi_1-\pi_2)e^{-\lambda} & \text{when } y=0\\ \pi_2+(1-\pi_1-\pi_2)\dfrac{\lambda^k e^{-\lambda}}{k!} & \text{when } y=k\\ (1-\pi_1-\pi_2)\dfrac{\lambda^y e^{-\lambda}}{y!} & \text{when } y\geq 1, y\neq k.\end{cases} \tag{1}$$

where $0<\pi_1<\pi_1+\pi_2<1$ and $\lambda>0$. We will refer to this distribution (1) as the zero and k inflated Poisson (ZkIP) distribution. The moment generating function of the ZkIP distribution is

$$M_Y(t)=E(e^{tY})=\pi_1+\pi_2\,e^{tk}+\pi_3\,e^{\lambda(e^t-1)}$$

and the probability generating function is

$$G_Y(z)=E(z^Y)=\pi_1+\pi_2\,z^k+\pi_3\,e^{\lambda(z-1)}.$$

where $\pi_3=(1-\pi_1-\pi_2)$. These functions could be used to show that the mean and variance of the ZkIP distribution are

$$\begin{aligned}E(Y) &= k\,\pi_2+\pi_3\,\lambda\\ Var(Y) &= k^2\pi_2(1-\pi_2)+\pi_3\lambda(1+\pi_1\lambda+\pi_2\lambda-2k\pi_2).\end{aligned}$$

The ZkIP distribution is essentially a mixture of Poisson and two degenerate distributions at zero and k. It can be elucidated as follows. Consider a latent variable

$\mathbf{z} = (z_1, z_2, z_3)$ which is distributed as multinomial with parameters $(1, \pi_1, \pi_2, \pi_3)$. Note that $\mathbf{z}$ takes values (1, 0, 0) with probability π_1; (0, 1, 0) with probability π_2 and (0, 0, 1) with probability π_3. That is,

$$P(\mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{if } z_1 = 1, z_2 = 0, z_3 = 0 \\ \pi_2 & \text{if } z_1 = 0, z_2 = 1, z_3 = 0 \\ \pi_3 & \text{if } z_1 = 0, z_2 = 0, z_3 = 1. \end{cases} \tag{2}$$

Further, let us assume the conditional distribution of $Y|\mathbf{z}$ is

$$P(Y = y|\mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} 1 & \text{for } z_1 = 1, y = 0 \\ 1 & \text{for } z_2 = 1, y = k \\ \dfrac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \ldots. \end{cases} \tag{3}$$

Thus, the joint distribution of (Y, $\mathbf{z}$), which is obtained by multiplying (2) and (3) is

$$P(Y = y,\, \mathbf{z} = (z_1, z_2, z_3)) = \begin{cases} \pi_1 & \text{for } z_1 = 1, y = 0 \\ \pi_2 & \text{for } z_2 = 1, y = k \\ \pi_3 \dfrac{\lambda^y e^{-\lambda}}{y!} & \text{for } z_3 = 1, y = 0, 1, \ldots. \end{cases} \tag{4}$$

The marginal of Y can now be obtained from (4) summing over the three possible values of $\mathbf{z}$. Thus we get

$$\begin{aligned} P(Y = 0) &= P(Y = 0,\, z_1 = 1) + P(Y = 0,\, z_2 = 1) + P(Y = 0,\, z_3 = 1) \\ &= \pi_1 + \pi_3 e^{-\lambda} \\ P(Y = k) &= P(Y = k,\, z_1 = 1) + P(Y = k,\, z_2 = 1) + P(Y = k,\, z_3 = 1) \\ &= \pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!} \end{aligned}$$

and

$$\begin{aligned} P(Y=y) &= P(Y=y,\, z_1=1) + P(Y=y,\, z_2=1) + P(Y=y,\, z_3=1) \\ &= \pi_3 \frac{\lambda^y e^{-\lambda}}{y!}, \quad \text{for} \quad y \geq 1 \;\; y \neq k. \end{aligned}$$

which is equivalent to the ZkIP distribution (1). Further, the posterior distribution $P(\mathbf{z}|Y) = P(\mathbf{z})P(Y|\mathbf{z})/P(Y)$ can be summarized in Table 1 below. Later, we will use the displayed conditional probabilities in this table to develop EM algorithm for estimation of the ZkIP parameters.

Table 1. $P(\mathbf{z} = (z_1, z_2, z_3)|Y = y)$ of ZkIP

$\mathbf{z} = (z_1, z_2, z_3)$	$y = 0$	$y = k$	$y \neq 0, k$
$z_1 = 1$	$\frac{\pi_1}{\pi_1 + \pi_3 p_0}$	0	0
$z_2 = 1$	0	$\frac{\pi_2}{\pi_2 + \pi_3 p_k}$	0
$z_3 = 1$	$\frac{\pi_3 p_0}{\pi_1 + \pi_3 p_0}$	$\frac{\pi_3 p_k}{\pi_2 + \pi_3 p_k}$	1

NOTE: The sum of entries in any column is one.

2.3 LIKELIHOOD FOR GROUPED DATA

Suppose that we have a vector $\mathbf{y} = (y_1, y_2, ..., y_n)$ consisting of a random sample of n observations from the ZkIP distribution. The frequency distribution of the sample can be organized in a table as

j	0	1	...	k	...	K	Total
Observed frequency	n_0	n_1	...	n_k	...	n_K	n

where $n_j = \#$ of y_i's that are equal to j, and $K = \max\{y_i\}$. If the observations are truly from the ZkIP distribution, the values of n_0 and n_k will be large compared to the rest of the frequencies. The vector of observed frequencies $(n_0, n_1, \ldots, n_K)$ can be regarded as incomplete data in the sense n_0 is actually $n_a + n_b$ and $n_k = n_c + n_d$,

where the unknown n_a and n_c are frequencies from degenerate distributions at 0 and k respectively.

The likelihood function of the observed frequencies from the ZkIP model is

$$\begin{aligned} L_{obs}(\pi_1, \pi_2, \lambda|\mathbf{y}) &\propto (\pi_1 + \pi_3 e^{-\lambda})^{n_0} \left(\pi_2 + \pi_3 \frac{\lambda^k e^{-\lambda}}{k!}\right)^{n_k} \prod_{j\neq 0,k}^{K} \left(\pi_3 \frac{\lambda^j e^{-\lambda}}{j!}\right)^{n_j} \\ &\propto (\pi_1 + \pi_3 p_0)^{n_0} (\pi_2 + \pi_3 p_k)^{n_k} \prod_{j\neq 0,k}^{K} (\pi_3 p_j)^{n_j}, \end{aligned} \tag{5}$$

where $\pi_3 = (1 - \pi_1 - \pi_2)$ and $p_j = \dfrac{e^{-\lambda}\lambda^j}{j!}$ for $j \geq 0$.

Note that when $\pi_2 = 0$ the ZkIP reduces to ZIP. Thus the likelihood for the ZIP model is

$$L_{obs}(\pi_1, \lambda|\mathbf{y}) \propto (\pi_1 + (1 - \pi_1)e^{-\lambda})^{n_0} \prod_{j\neq 0}^{K} \left((1 - \pi_2)\frac{\lambda^j e^{-\lambda}}{j!}\right)^{n_j}.$$

And if $\pi_1 = \pi_2 = 0$, the likelihood (5) becomes the likelihood function of the Poisson distribution

$$L_{obs}(\lambda|\mathbf{y}) = \prod_{j=0}^{K} \left(\frac{\lambda^j e^{-\lambda}}{j!}\right)^{n_j}.$$

2.4 ESTIMATION OF ZKIP PARAMETERS

The ZkIP is a three parameter distribution. Based on random sample of n observations, these three unknown parameters π_1, π_2 and λ can be estimated using the popular methods of estimation, the method of moments, and the maximum likelihood (ML). However, the method of moments procedure has been shown, (McLachlan and Peel, 2000), to yield inefficient estimates for finite mixture models (FMM) and it is expected to be the case for ZkIP distribution as well. Therefore, we will not pursue method of moments procedure in this dissertation. As the name suggests, the ML method involves maximizing the likelihood function. It is a common practice to deal with the loglikelihood instead of the likelihood, and the ML estimates are

obtained equating to zero the first order partial derivatives with respect to the unknown parameters. We also need to ensure that the Hessian matrix of the negative loglikelihood at the solution is positive definite.

Another popular alternative method for parameter estimation is the Expectation Maximum (EM) likelihood algorithm. Dempster et al. (1977) illustrated the use of the EM algorithm to obtain the maximum likelihood estimates for simple mixture models. Since then the EM method has become a popular method for parameter estimation in finite mixture models. The EM is an iterative method and has advantages over the ML method. An advantage is that it often uses closed form equations in the iterative process.

2.4.1 SCORE EQUATIONS FOR ML ESTIMATION

In this section, we provide details of the maximum likelihood estimation of the three π_1, π_2 and λ, unknown parameters of the ZkIP distribution. The maximum likelihood estimates are obtained by maximizing the log of the likelihood function. Using (5) and taking log on both sides we get the loglikelihood ℓ_{obs} of the observed data as

$$\ell_{obs}(\pi_1, \pi_2, \lambda|\mathbf{y}) \quad \propto \quad n_0 \log(\pi_1 + \pi_3 p_0) + n_k \log(\pi_2 + \pi_3 p_k) + \sum_{j \neq 0,k}^{K} n_j \ (\log \pi_3 + \log p_j)$$

or

$$\begin{aligned} \ell_{obs}(\pi_1, \pi_2, \lambda|\mathbf{y}) \quad \propto \quad & n_0 \log(\pi_1 + \pi_3 p_0) + n_k \log(\pi_2 + \pi_3 p_k) \\ & + (n - n_0 - n_k)(\log \pi_3 - \lambda) + \log(\lambda) \sum_{j \neq 0,k}^{K} (j \ n_j). \end{aligned} \quad (6)$$

We obtain the score equations for the ML estimation by taking the partial derivatives of (6) with respect to the three unknown parameters. The partial derivative with respect to π_1,

$$\frac{\partial \ell_{obs}(\pi_1, \pi_2, \lambda)}{\partial \pi_1} = 0$$

simplifies to

$$\frac{n_0(1-p_0)}{\pi_1 + \pi_3 p_0} = \frac{n - n_0 - n_k}{\pi_3} + \frac{n_k p_k}{\pi_2 + \pi_3 p_k}. \tag{7}$$

Similarly, the equation

$$\frac{\partial \ell_{obs}(\pi_1, \pi_2, \lambda)}{\partial \pi_2} = 0$$

can be seen to be equivalent to

$$\frac{n_k(1-p_k)}{\pi_2 + \pi_3 p_k} = \frac{n - n_0 - n_k}{\pi_3} + \frac{n_0 p_0}{\pi_1 + \pi_3 p_0}. \tag{8}$$

Finally, the partial derivative

$$\frac{\partial \ell_{obs}(\pi_1, \pi_2, \lambda)}{\partial \lambda} = 0$$

reduces to

$$\frac{1}{\lambda}\sum_{j \neq 0,k}^{K} (j n_j) - (n - n_0 - n_k) = \frac{n_0 \pi_3 p_0}{\pi_1 + \pi_3 p_0} - \frac{n_k \pi_3 p_k}{\pi_2 + \pi_3 p_k}\left(\frac{k}{\lambda} - 1\right). \tag{9}$$

The ML estimates of the three parameters π_1, π_2, λ are obtained by solving the three score equations (7), (8), and (9) simultaneously. The ML estimates can also be obtained directly by minimizing the negative of the loglikelihood function (6) using the optimization methods that are available in statistical software R and SAS. However, these optimizations routines may have convergence problems and may fail to yield the estimates.

2.4.2 ALTERNATIVE METHOD OF ESTIMATION

An alternate and computationally simpler approach for parameter estimation is the EM method. As noted earlier, the Expectation Maximum (EM) likelihood

method was introduced by Dempster et al. (1977) in a seminal paper. The algorithm is a simple modification of the maximum likelihood, and has become a popular alternative for ML estimation in cases where data are missing or incomplete. Zhang et al. (2016) used the EM approach to study the ZOIP model for grouped data. We extend their approach to our ZkIP model.

The frequency vector $(n_0, n_1, \ldots, n_k, \ldots, n_K)$ is the "observed" data. It can be viewed as the "incomplete" data, in the sense $n_0 = n_a + n_b$ and $n_k = n_c + n_d$, where the data is missing the number n_b of zeros and the number n_d of k's that are from Poisson distribution are missing. Here n_a, n_c are the unknown number of observations from degenerates distributions at 0 and k respectively. Thus, the complete data vector including the missing frequencies is $(n_a, n_b, n_1, \ldots, n_c, n_d, \ldots, n_K)$. The likelihood function of this complete data vector is

$$L_{comp}(\pi_1, \pi_2, \lambda|\mathbf{y}) \propto {\pi_1}^{n_a}\ {\pi_2}^{n_c}\ {\pi_3}^{(n-n_a-n_c)}\ {p_0}^{n_b}\ {p_k}^{n_d}\ \prod_{j\neq 0,k}^{K} {p_j}^{n_j} \tag{10}$$

where $\pi_3 = (1 - \pi_1 - \pi_2)$ and $p_j = \dfrac{e^{-\lambda}\lambda^j}{j!}$ for $j \geq 0$.

Our interest is to maximum or minimize the negative of the log of the above likelihood. The loglikelihood, $\ell_{comp} = \log L_{comp}$, can be written as

$$\begin{aligned}
\ell_{comp}(\pi_1, \pi_2, \lambda|\mathbf{y}) \quad &\propto \quad n_a \log(\pi_1) + n_c \log(\pi_2) + (n - n_a - n_c) \log \pi_3 \\
&\quad\ + n_b \log p_0 + n_d \log p_k + \sum_{j\neq 0,k}^{K} n_j \log p_j \\
&\propto \quad n_a \log(\pi_1) + n_c \log(\pi_2) + (n - n_a - n_c) \log \pi_3 \\
&\quad\ - n_b\lambda + n_d(-\lambda + k \log \lambda) + \sum_{j\neq 0,k}^{K} n_j(-\lambda + j \log \lambda). \qquad (11)
\end{aligned}$$

Please note the frequencies n_a and n_c are unknown. The expectation step in the EM algorithm replaces these frequencies with their expected values. These expected values can be computed by the posterior probabilities given in Table 1. More specifically,

$$\widehat{n}_a = n_0\, E(z_1/y=0) \;=\; n_0\, P(z_1=1/y=0) = n_0\,\frac{\pi_1}{\pi_1+\pi_3\, p_0},$$

$$\widehat{n}_c = n_k\, E(z_2/y=k) \;=\; n_k\, P(z_2=1/y=k) = n_k\,\frac{\pi_2}{\pi_2+\pi_3 p_k}. \tag{12}$$

The maximization step or the M-step in the EM algorithm involves maximizing the loglikelihood (11) after substituting these estimates for n_a and n_c. However, this maximization is easy since the score equations have closed form solutions. Indeed equating partial derivatives of (11) to zero we get,

$$\frac{\partial \ell_{comp}(\pi_1,\pi_2,\lambda)}{\partial \pi_1} = 0 \iff \widehat{\pi}_1 = \frac{n_a(1-\pi_2)}{n-n_c}, \tag{13}$$

$$\frac{\partial \ell_{comp}(\pi_1,\pi_2,\lambda)}{\partial \pi_2} = 0 \iff \widehat{\pi}_2 = \frac{n_c(1-\pi_1)}{n-n_a}, \tag{14}$$

$$\frac{\partial \ell_{comp}(\pi_1,\pi_2,\lambda)}{\partial \lambda} = 0 \iff \widehat{\lambda} = \frac{\sum_{j=0}^{K} j n_j}{n-n_a-n_c}. \tag{15}$$

Thus we summarize the steps of the EM algorithm as follows:

1. Choose initial values of π_1^0, π_2^0 and λ^0 for π_1, π_2 and λ respectively.
2. E-step: Calculate $\widehat{n}_a$, $\widehat{n}_c$ using (12), and set $\widehat{n}_b = n_0 - \widehat{n}_a$ and $\widehat{n}_d = n_1 - \widehat{n}_c$.
3. M-step: Update the estimates of π_1, π_2 and λ using the formulas in (13), (14) and (15).
4. Iterate the E-step and M-step until the estimates $\widehat{\pi}_1$, $\widehat{\pi}_2$ and $\widehat{\lambda}$ converge.

We have developed an R code for this algorithm and use it in two data analysis examples in Section 2.7.

2.5 STANDARD ERRORS FOR THE PARAMETER ESTIMATES

In this section we will study on how to obtain the standard errors for the parameter estimates. It is well known that the ML estimates are asymptotically normal

with covariance matrix given by the inverse of the Fisher information matrix. We will derive expressions for the elements of the information matrix.

2.5.1 STANDARD ERRORS FOR ML ESTIMATES

Recall that the maximum likelihood estimates for the given data are obtained by minimizing the negative of the loglikelihood function ℓ_{obs} given in equation (6). Therefore the asymptotic standard errors of the ML estimates depend on the Fisher information matrix $\mathcal{I}_{obs}$ computed from the observed data, which is given by

$$\mathcal{I}_{obs} = \begin{bmatrix} -\frac{\partial^2 \ell_{obs}}{\partial \pi_1^2} & -\frac{\partial^2 \ell_{obs}}{\partial \pi_1 \partial \pi_2} & -\frac{\partial^2 \ell_{obs}}{\partial \pi_1 \partial \lambda} \\ -\frac{\partial^2 \ell_{obs}}{\partial \pi_2 \partial \pi_1} & -\frac{\partial^2 \ell_{obs}}{\partial \pi_2^2} & -\frac{\partial^2 \ell_{obs}}{\partial \pi_2 \partial \lambda} \\ -\frac{\partial^2 \ell_{obs}}{\partial \lambda \partial \pi_1} & -\frac{\partial^2 \ell_{obs}}{\partial \lambda \partial \pi_2} & -\frac{\partial^2 \ell_{obs}}{\partial \lambda^2} \end{bmatrix}. \tag{16}$$

The specific formulas for the elements of the above matrix are

$$\begin{aligned}
-\frac{\partial^2 \ell_{obs}}{\partial \pi_1^2} &= (n - n_0 - n_k)\left(\frac{1}{\pi_3}\right)^2 + n_0\left(\frac{1-p_0}{\pi_1 + \pi_3 p_0}\right)^2 + n_k\left(\frac{p_k}{\pi_2 + \pi_3 p_k}\right)^2 \\
-\frac{\partial^2 \ell_{obs}}{\partial \pi_1 \partial \pi_2} &= -\frac{\partial^2 \ell_{obs}}{\partial \pi_2 \partial \pi_1} = \frac{(n - n_0 - n_k)}{{\pi_3}^2} - \frac{n_0 p_0 (1-p_0)}{(\pi_1 + \pi_3 p_0)^2} - \frac{n_k p_k (1-p_k)}{(\pi_2 + \pi_3 p_k)^2} \\
-\frac{\partial^2 \ell_{obs}}{\partial \pi_1 \partial \lambda} &= -\frac{\partial^2 \ell_{obs}}{\partial \lambda \partial \pi_1} = \frac{-n_0(1-\pi_2)p_0}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k \pi_2 p_k}{(\pi_2 + \pi_3 p_k)^2}\left(\frac{k}{\lambda} - 1\right) \\
-\frac{\partial^2 \ell_{obs}}{\partial \pi_2^2} &= (n - n_0 - n_k)\left(\frac{1}{\pi_3}\right)^2 + n_0\left(\frac{p_0}{\pi_1 + \pi_3 p_0}\right)^2 + n_k\left(\frac{1-p_k}{\pi_2 + \pi_3 p_k}\right)^2 \\
-\frac{\partial^2 \ell_{obs}}{\partial \pi_2 \partial \lambda} &= -\frac{\partial^2 \ell_{obs}}{\partial \lambda \partial \pi_2} = \frac{-n_0 \pi_1 p_0}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k(1-\pi_1)p_k}{(\pi_2 + \pi_3 p_k)^2}\left(\frac{k}{\lambda} - 1\right) \\
-\frac{\partial^2 \ell_{obs}}{\partial \lambda^2} &= \frac{1}{\lambda^2}\sum_{j \neq 0,k}^{K} j n_j - \frac{n_0 \pi_1 \pi_3 p_0}{(\pi_1 + \pi_3 p_0)^2} + \frac{n_k \pi_3 p_k}{(\pi_2 + \pi_3 p_k)} \\
&\quad \left(\frac{k}{\lambda^2} - \frac{\pi_2}{\pi_2 + \pi_3 p_k}\left(\frac{k}{\lambda} - 1\right)^2\right).
\end{aligned}$$

The matrix $\mathcal{I}_{obs}$ is negative of the Hessian matrix of the loglikelihood of the observed counts. The square root of the diagonal elements of the inverse matrix $\mathcal{I}_{obs}^{-1}$ gives the standard errors for the ML estimates of π_1, π_2, and λ. Please note that our formulas generalize the results of Zhang et al. (2016) who have derived similar formulas for zero and one inflated Poisson (ZOIP) model.

2.5.2 STANDARD ERRORS FOR EM ESTIMATES

The optimization algorithms routinely output a numerically computed Hessian matrix for the functions that are being optimized. However, calculation of the standard errors will be more accurate if analytical expressions are available. To compute the standard errors of the estimates obtained by the EM algorithm, we follow the approach described by Louis (1982). The relation between the likelihood of the complete, observed and missing data is given

$$L_{comp}(\theta \,|\, \mathbf{y}, \mathbf{z}) \;\; = \;\; L_{obs}(\theta|\mathbf{y}) \; L_{miss}(\theta \,|\, (\mathbf{z}|\mathbf{y})) \tag{17}$$

where $\mathbf{y}$ and $\mathbf{z}$ stand for the observed and missing data respectively. From (17) taking logs we get

$$\ell_{comp}(\theta \,|\, \mathbf{y}, \mathbf{z}) \;\; = \;\; \ell_{obs}(\theta|\mathbf{y}) + \; \ell_{miss}(\theta \,|\, (\mathbf{z}|\mathbf{y})) \tag{18}$$

Taking second order partial derivatives, we can see that from equation (18) the information matrices for the complete, observed and missing data satisfy the following identity

$$\mathcal{I}_{comp} \;\; = \;\; \mathcal{I}_{obs} + \mathcal{I}_{missing}$$

or

$$\mathcal{I}_{obs} \;\; = \;\; \mathcal{I}_{comp} - \mathcal{I}_{miss}. \tag{19}$$

Since the right hand side of equation (19) depends on the missing data, Louis (1982) has suggested to take the expected value of the missing data given the observed.

This gives us the identity

$$\mathcal{I}_{obs} = E(\mathcal{I}_{obs}|\mathbf{y}) = E(\mathcal{I}_{comp}|\mathbf{y}) - E(\mathcal{I}_{miss}|\mathbf{y}). \tag{20}$$

In other words, Louis (1982) estimate of the observed information matrix is given by

$$\widehat{\mathcal{I}_{obs}} = E(\mathcal{I}_{comp}|\mathbf{y}) - E(\mathcal{I}_{miss}|\mathbf{y}). \tag{21}$$

Equation (21) could be used to find the standard errors of the EM estimates. Note that

$$\mathcal{I}_{comp} = \begin{bmatrix} -\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \lambda} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2} & -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \lambda} \\ -\frac{\partial^2 \ell_{comp}}{\partial \lambda \partial \pi_1} & -\frac{\partial^2 \ell_{comp}}{\partial \lambda \partial \pi_2} & -\frac{\partial^2 \ell_{comp}}{\partial \lambda^2} \end{bmatrix}. \tag{22}$$

From (11), the elements of the information matrix $\mathcal{I}_{comp}$ are

$$\begin{aligned} -\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2} &= \frac{n_a}{{\pi_1}^2} + \frac{(n - n_a - n_c)}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2} &= -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} = \frac{(n - n_a - n_c)}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2} &= \frac{n_c}{\pi_2^2} + \frac{(n - n_a - n_c)}{\pi_3^2} \\ -\frac{\partial^2 \ell_{comp}}{\partial \lambda^2} &= \frac{n_d k}{\lambda^2} + \frac{\sum_{j \neq 0,k}^{K} j\, n_j}{\lambda^2}. \end{aligned}$$

The other elements $-\partial^2 \ell_{comp}/\partial \pi_1 \partial \lambda$ and $-\partial^2 \ell_{comp}/\partial \pi_2 \partial \lambda$ are equal to zero. Since n_a and n_c are missing, we replace them by their expected values

$$E(n_a|n_0) = \frac{n_0 \pi_1}{\pi_1 + \pi_3 p_0} \qquad \text{and} \quad E(n_c|n_k) = \frac{n_k \pi_2}{\pi_2 + (\pi_3 p_k)}.$$

Thus the nonzero elements of $E(\mathcal{I}_{comp}|\mathbf{y}) = E(\mathcal{I}_{comp}|n_0, n_k)$ are

$$E\left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_1^2}\right] = \frac{n}{\pi_3^2} + \frac{n_0}{\pi_1(\pi_1 + \pi_3 p_0)} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)}$$

and

$$
\begin{aligned}
E\left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_1 \partial \pi_2}\right] &= -\frac{\partial^2 \ell_{comp}}{\partial \pi_2 \partial \pi_1} = \frac{n}{\pi_3^2} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)} \\
E\left[-\frac{\partial^2 \ell_{comp}}{\partial \pi_2^2}\right] &= \frac{n}{\pi_3^2} - \frac{n_0 \pi_1}{\pi_3^2(\pi_1 + \pi_3 p_0)} + \frac{n_k}{\pi_2(\pi_2 + \pi_3 p_k)} - \frac{n_k \pi_2}{\pi_3^2(\pi_2 + \pi_3 p_k)} \\
E\left[-\frac{\partial^2 \ell_{comp}}{\partial \lambda^2}\right] &= \frac{n_k k}{\lambda^2} - \frac{n_k k \pi_2}{\lambda^2(\pi_2 + \pi_3 p_k)} + \frac{1}{\lambda^2} \sum_{j \neq 0,k}^{K} j n_j.
\end{aligned}
$$

Next to compute the second term $E(\mathcal{I}_{miss}|\mathbf{y})$ in equation (19), we proceed as follows. The likelihood of the observed and complete data are given in (5), (10) respectively. Hence, the likelihood of the missing data is obtained taking the ratio of these likelihoods and it is given by

$$
\begin{aligned}
L_{miss}(\pi_1, \pi_2, \lambda|\mathbf{z}) \propto\ & {\pi_1}^{n_a}\ {\pi_2}^{n_c}\ (p_0\pi_3)^{n_b}\ (p_k\pi_3)^{n_d} \\
& \left(\frac{1}{\pi_1 + \pi_3 p_0}\right)^{n_0} \left(\frac{1}{\pi_2 + \pi_3 p_k}\right)^{n_k}.
\end{aligned}
$$

Thus, the loglikelihood of the missing data is

$$
\begin{aligned}
\ell_{miss}(\pi_1, \pi_2, \lambda|\mathbf{y}) \propto\ & n_a \log(\pi_1) + n_c \log(\pi_2) - n_0 \log(\pi_1 + \pi_3 p_0) \\
& - n_k \log(\pi_2 + \pi_3 p_k) + (n_b + n_d) \log(\pi_3) \\
& - (n_b + n_d)\lambda + (n_d k) \log(\lambda). \qquad (23)
\end{aligned}
$$

We can easily check that the first order partial derivatives are

$$
\begin{aligned}
\frac{\partial \ell_{miss}}{\partial \pi_1} &= \frac{n_a}{\pi_1} - n_0 \left(\frac{1 - p_0}{\pi_1 + \pi_3 p_0}\right) - \frac{n_b + n_d}{\pi_3} - \frac{n_k p_k}{\pi_2 + \pi_3 p_k} \\
\frac{\partial \ell_{miss}}{\partial \pi_2} &= \frac{n_c}{\pi_2} + n_0 \left(\frac{p_0}{\pi_1 + \pi_3 p_0}\right) - \frac{n_b + n_d}{\pi_3} - \frac{n_k(1 - p_k)}{\pi_2 + \pi_3 p_k} \\
\frac{\partial \ell_{miss}}{\partial \lambda} &= \frac{n_d k}{\lambda} - (n_b + n_d) + n_0 \left(\frac{\pi_3 p_0}{\pi_1 + \pi_3 p_0}\right) \\
&- \frac{n_k \pi_3 p_k}{\pi_2 + \pi_3 p_k} \left(\frac{k}{\lambda} - 1\right).
\end{aligned}
$$

and the negative of the second order partial derivatives are

$$\begin{aligned}
-\frac{\partial^2 \ell_{miss}}{\partial \pi_1^2} &= \frac{n_a}{\pi_1^2} - \frac{n_0(1-p_0)^2}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k p_k^2}{(\pi_2+\pi_3 p_k)^2} + \frac{(n_b+n_d)}{\pi_3^2} \\
-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \pi_2} &= \frac{n_0 p_0(1-p_0)}{(\pi_1+\pi_3 p_0)^2} + \frac{n_k p_k(1-p_k)}{(\pi_2+\pi_3 p_k)^2} + \frac{(n_b+n_d)}{\pi_3^2} \\
-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \lambda} &= \frac{n_0(1-\pi_2)p_0}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k \pi_2 p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right) \\
-\frac{\partial^2 \ell_{miss}}{\partial \pi_2^2} &= \frac{n_c}{\pi_2^2} - \frac{n_0 p_0^2}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k(1-p_k)^2}{(\pi_2+\pi_3 p_k)^2} + \frac{(n_b+n_d)}{\pi_3^2} \\
-\frac{\partial^2 \ell_{miss}}{\partial \pi_2 \partial \lambda} &= \frac{n_0 \pi_1 p_0}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k(1-\pi_1)p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right) \\
-\frac{\partial^2 \ell_{miss}}{\partial \lambda^2} &= \frac{n_0 \pi_1 \pi_3 p_0}{(\pi_1+\pi_3 p_0)^2} + \frac{n_k \pi_2 \pi_3 p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right)^2 \\
&\quad -\frac{k}{\lambda^2}\frac{n_k \pi_3 p_k}{(\pi_2+\pi_3 p_k)} + \frac{k n_d}{\lambda^2}.
\end{aligned}$$

Once again using the expected values

$$E(n_a|n_0) = \frac{n_0 \pi_1}{\pi_1+\pi_3 p_0} \quad \text{and} \quad E(n_c|n_k) = \frac{n_k \pi_2}{\pi_2+(\pi_3 p_k)},$$

$$E(n_b|n_0) = \frac{n_0 \pi_3 p_0}{\pi_1+\pi_3 p_0} \quad \text{and} \quad E(n_d|n_k) = \frac{n_k \pi_3 p_k}{\pi_2+(\pi_3 p_k)},$$

we get the elements of $E(\mathcal{I}_{miss}|\mathbf{y}) = E(\mathcal{I}_{miss}|n_0, n_k)$ as follows

$$\begin{aligned}
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1^2}\right] &= \frac{n_0}{\pi_1(\pi_1+\pi_3 p_0)} - \frac{n_0(1-p_0)^2}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k p_k^2}{(\pi_2+\pi_3 p_k)^2} \\
&\quad + \frac{n_0 p_0}{\pi_3(\pi_1+\pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2+\pi_3 p_k)} \\
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \pi_2}\right] &= \frac{n_0 p_0(1-p_0)}{(\pi_1+\pi_3 p_0)^2} + \frac{n_k p_k(1-p_k)}{(\pi_2+\pi_3 p_k)^2} \\
&\quad + \frac{n_0 p_0}{\pi_3(\pi_1+\pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2+\pi_3 p_k)}
\end{aligned}$$

and

$$
\begin{aligned}
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_1 \partial \lambda}\right] &= \frac{n_0(1-\pi_2)p_0}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k \pi_2 p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right) \\
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_2^2}\right] &= \frac{n_k}{\pi_2(\pi_2+\pi_3 p_k)} - \frac{n_0 p_0^2}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k(1-p_k)^2}{(\pi_2+\pi_3 p_k)^2} \\
&\quad + \frac{n_0 p_0}{\pi_3(\pi_1+\pi_3 p_0)} + \frac{n_k p_k}{\pi_3(\pi_2+\pi_3 p_k)} \\
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \pi_2 \partial \lambda}\right] &= \frac{n_0 \pi_1 p_0}{(\pi_1+\pi_3 p_0)^2} - \frac{n_k(1-\pi_1)p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right) \\
E\left[-\frac{\partial^2 \ell_{miss}}{\partial \lambda^2}\right] &= \frac{n_0 \pi_1 \pi_3 p_0}{(\pi_1+\pi_3 p_0)^2} + \frac{n_k \pi_2 \pi_3 p_k}{(\pi_2+\pi_3 p_k)^2}\left(\frac{k}{\lambda}-1\right)^2 .
\end{aligned}
$$

The remaining elements follow by symmetry.

2.6 HYPOTHESIS TESTING AND MODEL SELECTION

In statistical inference, estimation of the parameters is usually followed by testing significance of the parameters and selection of the best model for the data. Hence, in this section we discuss the hypothesis testing to see whether there is a significant inflation at k and at zero. In other words, whether ZkIP is significantly fits the data better than the ZIP or the simpler Poisson model.

There are various criterions to select the best model. We use the Akaike Information Criterion (AIC) and likelihood-ratio method to arrive at the best model that fits the data. These details will be illustrated with a couple of real life data analysis.

2.6.1 HYPOTHESIS TESTING

Recall that the ZkIP is a three parameter distribution. The parameters π_1 and π_2 represent the proportion of observations that come from degenerate distributions and the parameter λ represents the mean of the Poisson distribution. Let $\widehat{\theta} = (\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\lambda})$ denote the ML or EM estimates of these parameters. Assume the true value $\theta^0 = (\pi_1^0, \pi_2^0, \lambda^0)$ is in the interior of the parameter space, that is, $0 < \pi_1^0 + \pi_2^0 < 1$ and $\lambda^0 > 0$. Under usual regularity conditions $\widehat{\theta}$ is asymptotically normal with mean θ^0 and covariance matrix given by $(\widehat{\mathcal{I}_{obs}})^{-1}$. We can use this result to construct a Wald's test for testing the hypotheses that a specified proportion $0 < \pi_2^0 < 1$ of observations

come from a degenerate distribution at k or a specified proportion $0 < \pi_1^0 < 1$ come from the degenerate distribution at zero. Similarly the hypothesis $H_0 : \lambda = \lambda_0 > 0$ could be tested for significance using the Wald's test.

The FMM and Countreg procedures in SAS use the parameters $\gamma = \log(\pi_1/\pi_3)$ and $\delta = \log(\pi_2/\pi_3)$ and test for the hypothesis $H_0 : (\gamma, \delta) = (0, 0)$. This hypothesis is equivalent to testing $H_0 : (\pi_1, \pi_2) = (\pi_1^0 = 1/3, \pi_2^0 = 1/3)$, which we could do using Wald's test because $\pi_1^0 = 1/3$ and $\pi_2^0 = 1/3$ are values in the interior of the parameter space.

As we discussed in Section 2.2, the ZkIP, ZIP and the Poisson model form a group of three nested models in the sense Poisson is a special case of ZIP which is a special case of ZkIP. Thus one could use the Likelihood ratio test (LRT) to test the significance of the nested models, that is, whether the ZkIP model could be replaced by the ZIP model or whether the ZIP model could be replaced by the Poisson model. We need to test the null hypothesis $H_0 : \pi_2 = 0$ to see whether there is a significant or insignificant inflated frequency at count k. That is acceptance of the null hypothesis would imply we could replace ZkIP model with the ZIP model. Since $0 \leq \pi_2 \leq 1$, the null hypothesis $H_0 : \pi_2 = 0$ corresponds to testing a parameter value on the boundary. And therefore standard asymptotic theory for the likelihood ratio statistic is not applicable. And the asymptotic distribution of the likelihood ratio statistic in not a χ^2 distribution but it is a mixture of χ^2 distributions (Chant, 1974; Shapiro, 1985). In the next section we will sketch a general proof to show indeed the asymptotic distribution of the LRT statistic is a mixture of chi-square distributions.

2.6.2 ASYMPTOTIC DISTRIBUTION ON THE BOUNDARY

This section is independent of the rest of the dissertation and contains a general result on the asymptotic distribution of the likelihood ratio statistic that is applicable for mixture of any two distributions. Suppose we have a random sample of $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of n observations from the mixture distribution $\pi\, f(y, \eta) + (1 - \pi)\, f(y, \theta)$ where $f(y, \eta)$ and $f(y,\, \theta)$ are two distinct univariate densities and $0 \leq \pi \leq 1$. We assume that η and θ are known and are interested in testing

the null hypothesis $H_0 : \pi = 0$ versus $H_1 : \pi > 0$. The likelihood function is

$$\begin{aligned} L(\pi|\mathbf{y}) &= \prod_{i=1}^{n}(\pi f(y_i,\eta) + (1-\pi)f(y_i,\theta)) \\ &= \prod_{i=1}^{n} f(y_i,\theta)\left(\pi\frac{f(y_i,\eta)}{f(y_i,\theta)} + (1-\pi)\right) \\ &= \prod_{i=1}^{n} f(y_i,\theta)\left(1+\pi\left(\frac{f(y_i,\eta)}{f(y_i,\theta)} - 1\right)\right) \\ &= \prod_{i=1}^{n} f(y_i,\theta)(1+\pi u_i) \end{aligned}$$

where $u_i = f(y_i,\eta)/f(y_i,\theta) - 1$. The loglikelihood is

$$\begin{aligned} \ell(\pi|\mathbf{y}) = \log L(\pi|\mathbf{y}) &= \sum_{i=1}^{n}\log f(y_i,\theta) + \sum_{i=1}^{n}\log(1+\pi u_i) \\ &= \ell(\pi=0|\mathbf{y}) + g(\pi), \end{aligned} \tag{24}$$

where $\ell(\pi = 0|\mathbf{y}) = \sum_{i=1}^{n}\log f(y_i,\theta)$ and $g(\pi) = \sum_{i=1}^{n}\log(1+\pi u_i)$. Under $H_0 : \pi = 0$, we have $E(u_i) = 0$ and let $V(u_i) = \sigma^2$. To derive the likelihood ratio test statistic we need to maximize (24) and this amounts to maximizing the function

$$\begin{aligned} g(\pi) &= \sum_{i=1}^{n}\log(1+\pi u_i) \approx \sum_{i=1}^{n}\left(\pi u_i - \frac{\pi^2 {u_i}^2}{2}\right) \\ &= \pi\sum_{i=1}^{n} u_i - \pi^2\sum_{i=1}^{n}\frac{{u_i}^2}{2} \end{aligned} \tag{25}$$

We have used the approximation $\log(1+u) \approx u - u^2/2$ for small u in equation (25). It is easy to check that $g(\pi)$ has a maximum at $(\sum_{i=1}^{n} u_i)/(\sum_{i=1}^{n} {u_i}^2)$. Since $\pi \geq 0$, the feasible point of maximum for $g(\pi)$ or equivalently for the loglikelihood is given by

$$\widehat{\pi} = \begin{cases} \dfrac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} {u_i}^2} & \text{when} \quad \sum_{i=1}^{n} u_i > 0 \\ \\ 0 & \text{when} \quad \sum_{i=1}^{n} u_i < 0. \end{cases}$$

Substituting $\widehat{\pi}$ in (25) we get

$$g(\widehat{\pi}) = \begin{cases} \dfrac{1}{2}\dfrac{n\overline{u}^2}{\frac{1}{n}\sum_{i=1}^{n} {u_i}^2} & \text{when} \quad \overline{u} > 0 \\ \\ 0 & \text{when} \quad \overline{u} < 0 \end{cases} \tag{26}$$

where $\overline{u} = (\sum_{i=1}^{n} u_i)/n$. Thus the maximum of the loglikelihood is

$$\ell(\widehat{\pi}|\mathbf{y}) = \ell(\pi = 0|\mathbf{y}) + g(\widehat{\pi}).$$

Therefore log of the likelihood ratio statistic, Λ, for the hypothesis $H_0 : \pi = 0$ is

$$\log \Lambda = \ell(\pi = 0|\mathbf{y}) - \ell(\widehat{\pi}|\mathbf{y}) = -g(\widehat{\pi}),$$

where $g(\widehat{\pi})$ is given by (26). Alternatively,

$$-2 \log \Lambda = 2\, g(\widehat{\pi}) = \frac{\sigma^2}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} {u_i}^2}\ \frac{n\overline{u}^2}{\sigma^2}\ I(\overline{u} > 0)$$

where $I(\cdot)$ is the indicator function. By the central limit theorem $\sqrt{n}\,\overline{u}/\sigma$ converges to the standard normal variable Z as $n \to \infty$. Therefore $P(\overline{u} > 0)$ converges $P(Z > 0) = 1/2$. Also, $n\overline{u}^2/\sigma^2$ converges in distribution to Z^2 which is distributed as χ^2_1 as $n \to \infty$. Further by the law of large numbers $\frac{1}{n}\sum_{i=1}^{n} u_i^2$ converges to σ^2 as $n \to \infty$. Putting all these together we can see that under $H_0 : \pi = 0$ we have as $n \to \infty$,

$$-2 \log \Lambda \xrightarrow{d} \frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1.$$

The above asymptotic distribution is useful to calculate the asymptotic significance level for testing $H_0 : \pi_2 = 0$, that is, ZkIP could be reduced to the ZIP model or for testing $H_0 : \pi_1 = 0$, that is, ZIP could be further reduced to the ordinary Poisson model.

2.6.3 GOODNESS OF FIT

For count data the most commonly used statistic for testing the goodness-of-fit test is the Pearson chi-square statistic $\chi^2 = \sum_{i=1}^{c}(o_i - e_i)^2/e_i$, where o_i is observed frequency and e_i is the expected frequency of the of the ith category, and c is the total number of categories. Asymptotically, the χ^2 statistic follows a chi-square distribution with $(c-1)$ degrees of freedom. The test is not the best when there are inflated frequencies. An alternate and a simple measure for checking the goodness-of-fit among competing models is the Absolute Error (ABE), which is defined as

$$\text{ABE} \;=\; \sum_{i=1}^{c} |o_i - e_i|.$$

Clearly, the model that has minimum ABE has the least deviation between the observed and expected frequencies. Hence, the model with minimum ABE fits data the best.

2.6.4 MODEL SELECTION

We can use several criteria for selecting the appropriate model between the three competing models, Poisson, ZIP and ZkIP. One criteria that we outlined in the previous section is the ABE. However, the ABE criteria tends to be very subjective. A popular criteria is the Akaike Information Criteria (AIC). The AIC was introduced by Akaike (1974) and it is calculated as $-2\,\ell + 2m$, where ℓ is the maximum value of the loglikelihood and m is the number of parameters for the model under consideration. The loglikelihood tends to increase as we move from a simpler model to a complex one. The constant $2\,m$ penalizes the complex model since it will have more parameters than the simple model. This avoids over fitting the model for the data. We select the model that has the minimum AIC as the best model.

2.7 EXAMPLES

In this section we illustrate the results presented in Sections 2.4 and 2.5 on two real life data. These data examples were obtained from the National Health Interview Survey (NHIS) conducted by the National Center for Health Sciences (NCHS). Since 1957, NCHS has been collecting and archiving data on US residents. The data are collected annually on various health issues including immunizations, depression, hepatitis, cancer, use of tobacco and other variables related to health. For our illustration we took a subset of data that was collected in year 2015. In both examples we fit zero and k inflated Poisson (ZkIP) model and compare it to the zero inflated Poisson (ZIP) and ordinary Poisson models. The first example illustrates a ZkIP model with inflations at 0 and $k = 6$, while the second example has inflations at 0 and $k = 1$. The latter model is also known as zero and one inflated Poisson (ZOIP) model.

2.7.1 PAP SMEAR DATA

Cervical cancer is of a major concern for the female population. A common preventive and early detection screening procedure for cervical cancer is the pap smear test. In this example, we looked at the number of pap smear tests a female took in last six years for females aged more than 18 years. The count variable represents responses to two questions in the survey: (1) *Have you ever had a Pap smear or Pap test?* and (2) *How many Pap tests have you had in the last 6 years?*. If the reply to the first question is a '*No*' then the number of tests done is reported zero, while if the reply is a '*Yes*' then the number of tests done is same as the reply to the second question.

There were a total of 33672 females interviewed in the survey, out of which about 3.5% choose not to answer or their response was not recorded. We performed a list wise deletion to clean the data and ended up with a data set consisting of $n = 12014$ cases. The mean number of the pap smear tests for this clean data is 3.40 and the variance is 5.25. The percentage (count) of females who never took a pap smear test was 15.68% (1884) and the percentage (count) of females who had one pap smear each year for a total of six in the last six years was 29.17% (3504). The observed

frequencies are presented in Table 3. Clearly the proportions of zero and six in the data set are inflated and both these proportions are more than what we would expect under a Poisson model. Thus an appropriate model for this data is the zero and six inflated Poisson model or the ZkIP model with $k = 6$.

Table 2. Results for pap smear grouped data

Parameter	ZkIP	ZIP	Poisson
$\widehat{\lambda}$	2.9820	3.9495	3.3957
	(0.0241)	(0.0197)	(0.0166)
$\widehat{\pi}_1$	0.1257	0.1403	–
	(0.0026)	(0.0022)	
$\widehat{\pi}_2$	0.2613	–	–
	(0.0018)		
$\log L_{obs}$	-23261.63	-26101.09	-28030.58
AIC	46529.26	52206.17	56063.00

NOTE: The EM standard errors are given in parenthesis.

The parameter estimates and standard errors for the three models ZkIP, ZIP and Poisson for the pap smear data are displayed in Table 2. Further, using the estimates we obtain the expected counts under the three probability models. The observed and expected frequencies are in Table 3 and a plot of these frequencies is in Figure 1.

Table 3. Frequencies for pap smear grouped data

Count	Observed	ZkIP	ZIP	Poisson
0	1884	1883.47	1884.53	402.71
1	1417	1113.23	785.81	1367.44
2	1362	1659.83	1551.79	2321.65
3	1536	1649.89	2042.96	2627.81
4	1115	1230.00	2017.19	2230.76
5	905	733.57	1593.39	1514.97
6	3504	3503.85	1048.86	857.38
> 6	291	155.32	591.79	415.90
ABE		1138.28	5674.97	8079.58
χ^2		313.54	7256.70	15311.83

Table 3 shows that the Poisson model terribly under estimates the frequencies at zero and six. The ZIP model is able to capture the inflation at zero but fails to capture the inflation at six. We can see this more clearly in Figure 1. Finally, it is the ZkIP model that is able to capture both the spikes at zero and at six. It has the

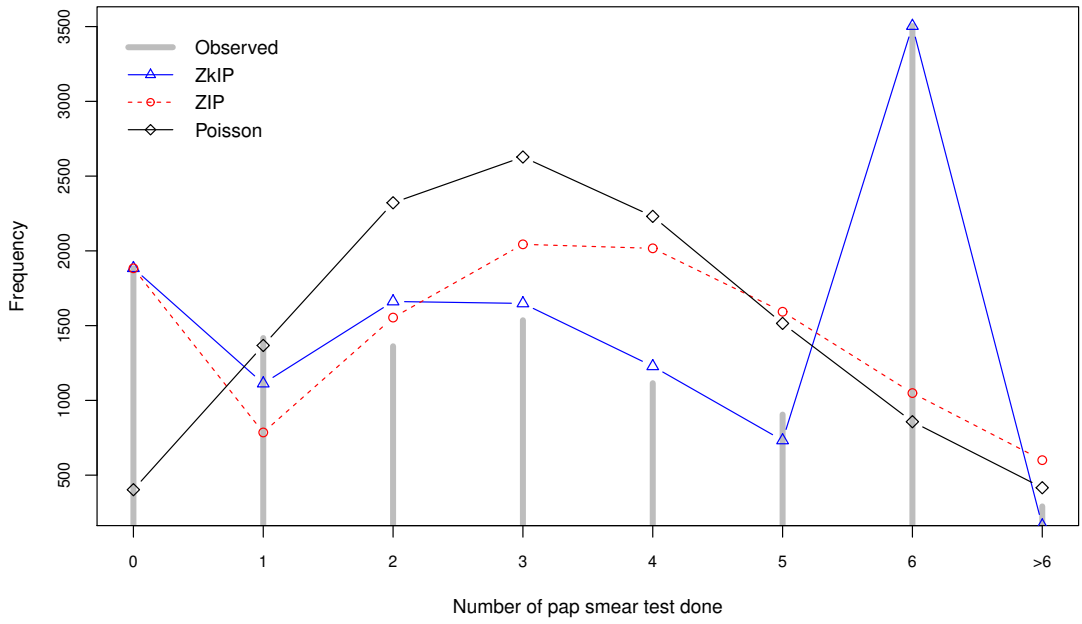


Figure 1. Observed and expected frequencies for pap smear grouped data.

minimum ABE and fits the data best among the three models.

Table 2 displays the AIC values for the three models. These values are 56063 for Poisson, 52206.17 for ZIP and, 46529.96 the smallest AIC for ZkIP. Once again ZkIP is the preferred model according to the AIC criterion. The LRT statistic for the hypothesis $H_0 : \pi_2 = 0$ versus $H_1 : \pi_2 > 0$ is given by $-2 \log \Lambda = 5678.92$, which is highly significant, and confirms that ZkIP could not be replaced by ZIP. Similarly, the LRT also shows that ZIP is superior to the Poisson model. Note that the estimates of π_1 and π_2 are 12.57% and 26.13% respectively, and these values are highly significant.

We further verify that the EM estimates of the ZkIP model do maximize the likelihood function. The negative of the loglikelihood function of the ZkIP model for $0 \leq \lambda \leq 5$ and $0 \leq \pi_2 \leq 0.4$ is plotted in the Figure 2. The Figure shows the EM estimates are close to $\lambda = 3$ and $\pi_2 = 0.2$. The results agree with the output in the Table 2. Hence, we conclude the ZIP model is a better count model then Poisson when there is inflation at zero. However, for inflation at zero and k the mixture

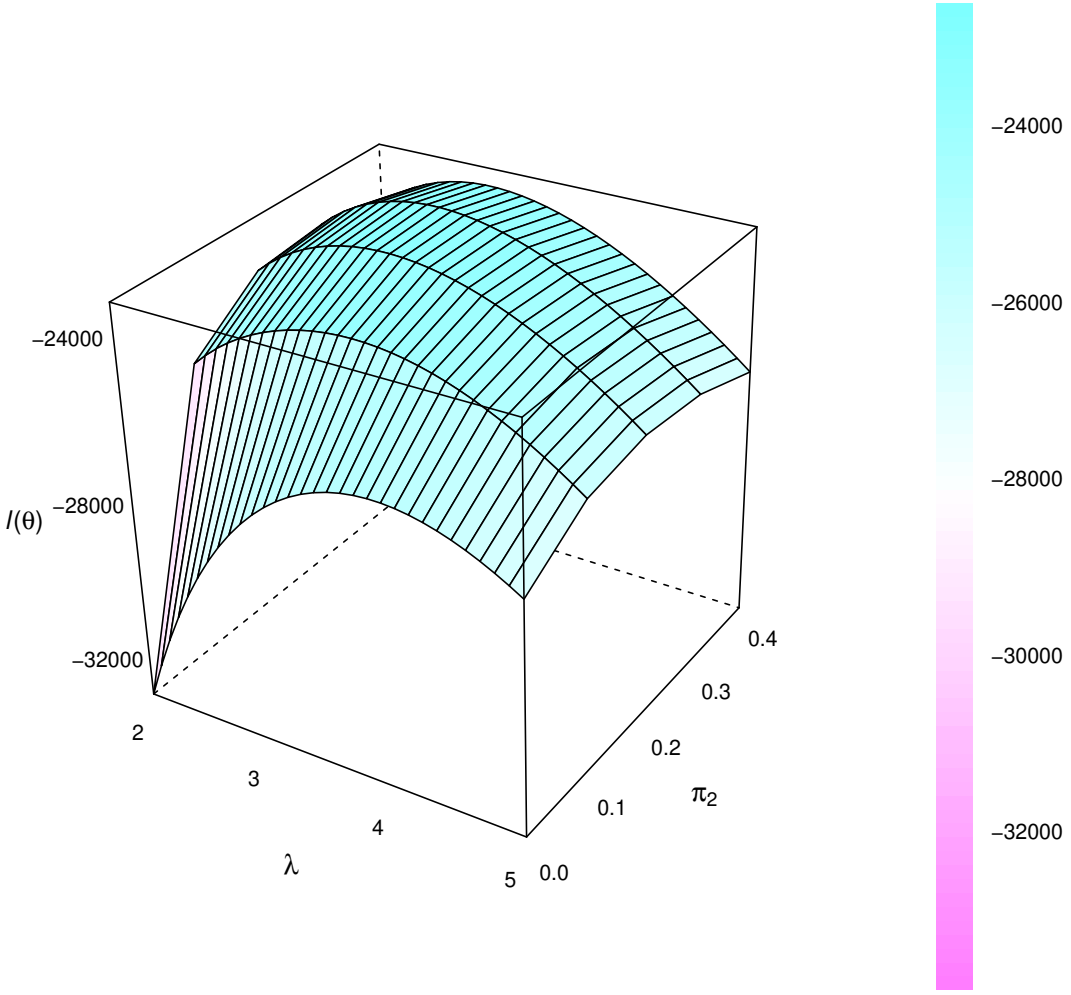


Figure 2. Loglikelihood function for grouped pap smear data for the ZkIP model.

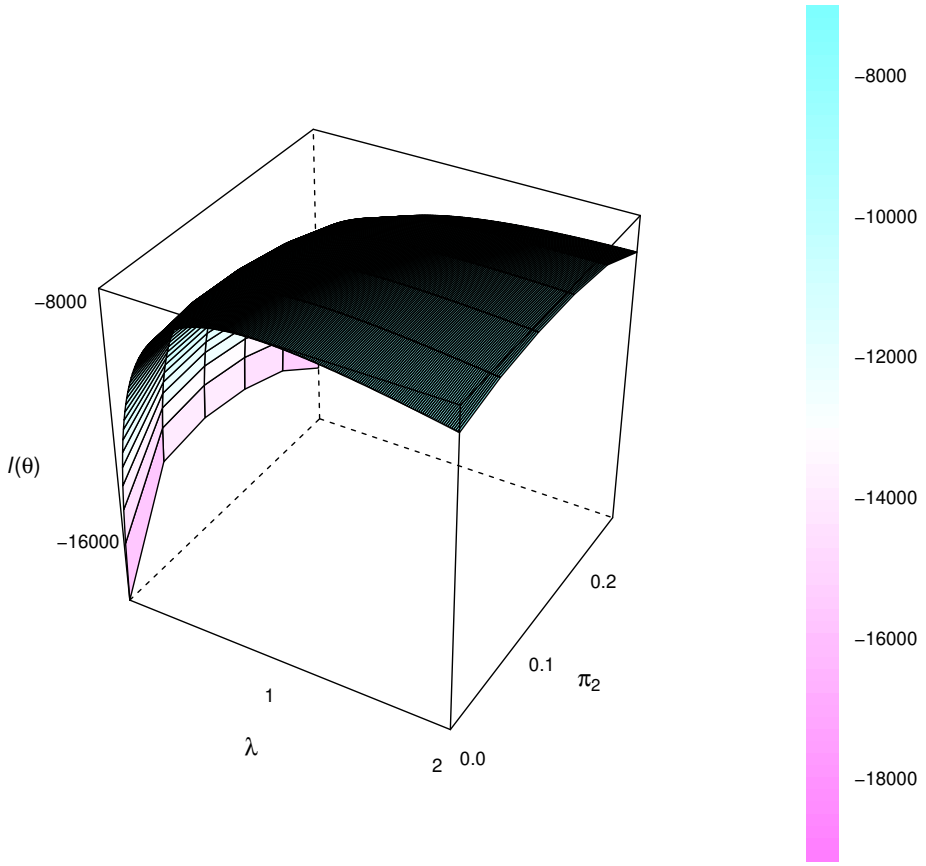


Figure 3. Loglikelihood function of ER data without covariate for the ZOIP model.

model zero and k inflated Poisson model gives a better fit.

2.7.2 EMERGENCY ROOM DATA

As a second example, we consider another count data that has inflated frequencies for two count values. This data set is also taken from the NHIS database for the year 2015. The data set consists of children less than 18 years old. The count variable is the number of visits a child made to an Emergency Room (ER) or an Emergency Department (ED) during the previous twelve months. We observe inflated number of zeros with a frequency of 10046, and account for 82% of the sample. The frequency and proportion of ones are 1466 and 12% respectively. For this data we fit and compare ZOIP, ZIP and the Poisson models.

The EM estimates and standard errors for the three models are displayed in Table 4. Using the AIC measure, the ZOIP model has the minimum AIC of 15502.02. Hence, ZOIP is the best model. The LRT statistics for $H_0 : \pi_1 = 0$ and $H_0 : \pi_2 = 0$ are respectively $-2\log\Lambda = 1181.66$ and $-2\log\Lambda = 28.46$. Both these are highly significant. Thus, according to LRT criterion ZOIP outperforms ZIP which in turn outperforms the Poisson model for this data. The loglikelihood function of the ZOIP model is plotted in Figure 3 for values of λ between 0 and 2 and values of π_2 between 0 and 0.3. Clearly the loglikelihood is a smooth function with a unique maximum.

Table 4. Results for ER grouped data

Parameter	ZOIP	ZIP	Poisson
$\widehat{\lambda}$	1.0618	0.8177	0.2607
	(0.0569)	(0.0252)	(0.0046)
$\widehat{\pi}_1$	0.7518	0.6811	–
	(0.0145)	(0.0091)	
$\widehat{\pi}_2$	0.0455	–	–
	(0.0352)		
$\log L_{obs}$	-7748.01	-7762.24	-8353.07
AIC	15502.02	15528.48	16708.00

NOTE: The EM standard errors are given in parenthesis.

For checking the goodness of fit we compared the observed and expected frequencies for the three models. As seen from Table 5, the ZOIP model has the smallest ABE and fits the data best. In conclusion as seen in Figure 4, the ZOIP is a clear winner for this data.

Table 5. Frequencies for ER grouped data

Count	Observed	ZOIP	ZIP	Poisson
0	10046	10046.06	10045.86	9417.61
1	1466	1465.93	1407.02	2455.53
2-3	548	483.02	575.24	320.13
4-5	92	170.96	156.78	27.82
6-7	37	45.38	32.05	1.81
8-9	12	9.64	5.24	0.09
10-12	13	1.71	0.71	0.00
> 12	9	0.26	0.08	0.00
ABE		174.85	184.05	1979.06
χ^2		417.45	1204.28	573117.10

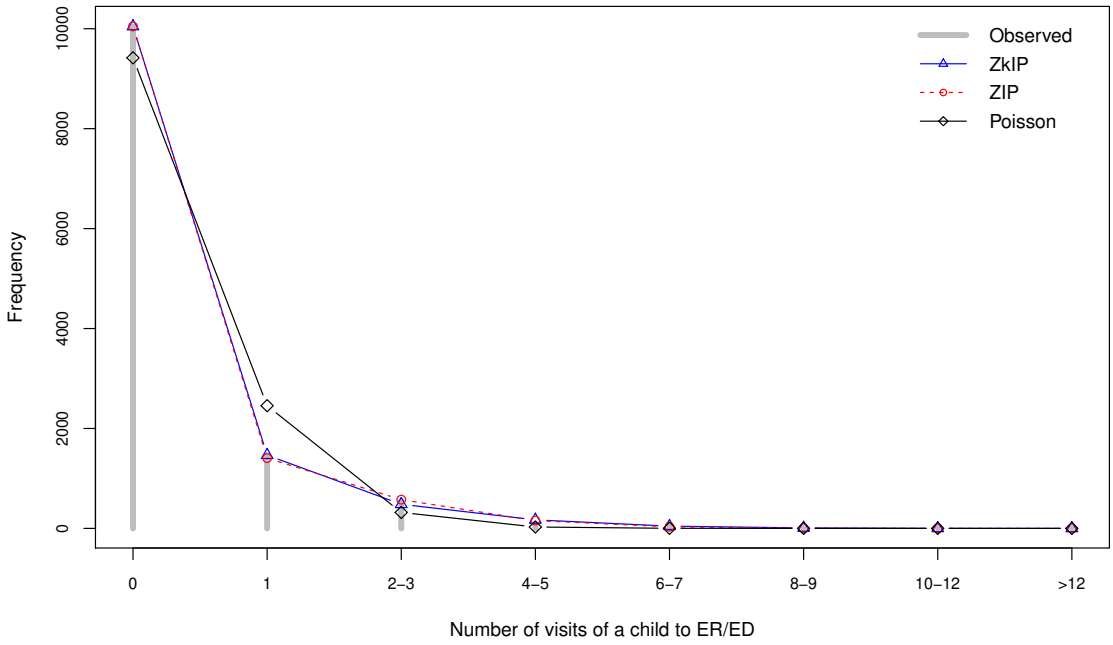


Figure 4. Observed and expected frequencies for ER data without covariate.

Hence, we conclude the ZOIP model fits better than ZIP and Poisson model to the count of number of times a child is taken to ER or ED.

CHAPTER 3

ZERO AND K INFLATED POISSON REGRESSION MODELS

3.1 INTRODUCTION

In Chapter 2, we have introduced the zero and k inflated Poisson model, and discussed estimation of the parameters using maximum likelihood and the expectation maximization algorithm. However, the discussion was limited to grouped data. In this chapter we assume besides the count responses the data also consists of covariate measurements on each subject. In this chapter, we consider the zero and k inflated Poisson regression models to study the dependence of the response variable on the covariates.

The outline of the chapter is as follows. We present the zero and k inflated Poisson regression model in Section 3.2. We describe the maximum likelihood and expectation maximum algorithm to estimate the regression parameters in Section 3.3. Computation of the standard errors for the regression estimates using the method described by Louis (1982) is presented in Section 3.4. Lastly, we illustrate our theory on two real life data sets in Section 3.5 including the identification of significant covariates. Finally, we compare the various Poisson models and find the best fit model using the AIC criterion.

3.2 ZKIP REGRESSION MODEL

In Section 2.2, we introduced and motivated the ZkIP distribution through a latent variable formulation. Recall, the probability density of the ZkIP distribution

is given by

$$P(Y=\mathsf{y}) = \begin{cases} \pi_1 + \pi_3\, p_0(\lambda) & \text{when} \quad y = 0 \\ \pi_2 + \pi_3\, p_k(\lambda) & \text{when} \quad y = k \\ \pi_3\, p_j(\lambda) & \text{when} \quad y \geq 1, y \neq k. \end{cases} \tag{27}$$

where $p_j(\lambda) = \lambda^y\, e^{-\lambda}/y!$ for $y \geq 0$ and $\pi_1 + \pi_2 + \pi_3 = 1$. Note that the ZkIP is a distribution with three parameters π_1, π_2 and λ. It is a mixture of three distributions. The first distribution is degenerate at zero with probability π_1 and the second is degenerate at k with probability π_2, and the third is the Poisson with mean λ and probability π_3. Clearly, the ZkIP distribution is reduced to ZIP when $\pi_2 = 0$ and Poisson distribution is a special case of ZIP for $\pi_1 = 0$. Some distributional properties of the ZkIP distribution are given in the Section 2.2.

Suppose that we have a vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of n independent count responses from a ZkIP distribution. Associated with each y_i, we assume a vector $\mathbf{x_i}^T = (x_{i1}, \ldots, x_{ip})$ of covariates has been observed. The layout of the observed data can be written in the following table where we assume the number of y_i's that are equal to 0 (or k) is high.

Observation	Response	Covariates
1	y_1	$x_{11} \quad \ldots \quad \ldots \quad x_{1p}$
2	y_2	$x_{21} \quad \ldots \quad \ldots \quad x_{2p}$
$\vdots$	$\vdots$	$\vdots \qquad\qquad\qquad \vdots$
i	y_i	$x_{i1} \quad \ldots \quad \ldots \quad x_{ip}$
$\vdots$	$\vdots$	$\vdots \qquad\qquad\qquad \vdots$
n	y_n	$x_{n1} \quad \ldots \quad \ldots \quad x_{np}$

From (27), the likelihood function of the observed data is

$$L_{obs}(\pi_1,\, \pi_2,\, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{i:y_i=0} (\pi_1 + \pi_3 p_{0i}(\lambda_i)) \prod_{i:y_i=k} (\pi_2 + \pi_3 p_{ki}(\lambda_i)) \prod_{i:y_i\neq 0,k} (\pi_3\, p_{yi}(\lambda_i)) \tag{28}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $p_{y_i(\lambda_i)} = e^{-\lambda_i}\lambda_i{}^{y_i}/y_i!$ for $y_i \geq 0$. To connect the parameters with the covariates we follow the standard generalized linear model (GLM) framework for the multinomial distribution. The three mixing distributions can be viewed as three nominal categories. Thus the probabilities of the three (degenerate(0), degenerate(k), Poisson) categories are π_1, π_2 and π_3 respectively. Following the GLM baseline category logits model for the multinomial we re-parametrize and set

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \gamma \quad \text{and} \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \delta. \tag{29}$$

Here we were treating the Poisson distribution as the baseline category and thus we have $(3-1) = 2$ equations for the other two categories. As in log-linear models, our ZkIP regression model assumes the Poisson parameter λ_i is a loglinear function of the covariates and it is given by

$$\log(\lambda_i) = \mathbf{x_i^T}\boldsymbol{\beta}.$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^T$ is a p dimensional unknown regression parameter. For simplicity we assume the parameters γ and δ are constants. The generalization where these two parameters are functions of the covariates is straight forward. Thus the parameters of our ZkIP regression model are $\boldsymbol{\beta}$, γ and δ, and we consider estimation of these parameters in the next section.

3.3 ESTIMATION OF THE REGRESSION PARAMETERS

In this section we will study methods for estimating the parameters of the ZkIP regression model. The two popular methods are the maximum likelihood (ML) and expectation maximization (EM) method. The ML technique involves optimizing the likelihood or the loglikelihood function with respect to the unknown parameters $\boldsymbol{\beta}$, γ and δ. Substituting the reparametrizations (29) in the likelihood function (28) we

get

$$\begin{aligned}\ell_{obs}(\boldsymbol{\beta}, \gamma, \delta) &= \log L_{obs}(\boldsymbol{\beta}, \gamma, \delta|\mathbf{y}) \\ &= \sum_{i:y_i=0} \log\left(e^{\gamma} + p_{0i}(\lambda_i)\right) + \sum_{i:y_i=k} \log\left(e^{\delta} + p_{ki}(\lambda_i)\right) \\ &\quad + \sum_{i:y_i\neq 0,k} \log(p_{yi}(\lambda_i)) - n\log(1 + e^{\gamma} + e^{\delta}) \end{aligned} \tag{30}$$

where $\log \lambda_i = x_i^T \boldsymbol{\beta}$. The ML estimates can be obtained maximizing the loglikelihood (30) directly with respect to the parameters or taking the partial derivatives and solving the three score equations (Lin and Tsai, 2012) given below,

$$\begin{aligned}\sum_{i:y_i=0} \frac{e^{\gamma}}{e^{\gamma} + p_{0i}(\lambda_i)} &= \frac{ne^{\gamma}}{(1 + e^{\gamma} + e^{\delta})} \\ \sum_{i:y_i=k} \frac{e^{\delta}}{e^{\delta} + p_{ki}(\lambda_i)} &= \frac{ne^{\delta}}{(1 + e^{\gamma} + e^{\delta})} \\ \sum_{i:y_i\neq 0,k} (y_i - \lambda_i)\mathbf{x_i} &= \sum_{i:y_i=0} \frac{\lambda_i\, p_{0i}(\lambda_i)}{e^{\gamma} + p_{0i}} \mathbf{x_i} - \sum_{i:y_i=k} \frac{(k - \lambda_i)\, p_{ki}(\lambda_i)}{e^{\delta} + p_{ki}(\lambda_i)} \mathbf{x_i}. \end{aligned} \tag{31}$$

These equations (31) can be solved iteratively using the Newton-Raphson method to obtain the ML estimates.

An alternative and popular method for parameter estimation is the expectation-maximization (EM) approach. The EM approach treats the observed data $\mathbf{y} = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})$ as part of a complete data that includes $\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2}, \ldots, \mathbf{z_n})$ which is regarded as missing. Here each $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ is a three component vector with probability distribution given by (2) and the conditional distribution of y_i given $\mathbf{z_i}$ is given by (3). Thus the joint distribution of the observed and missing is given by

$$P(y_i, \mathbf{z_i}) = \begin{cases} \pi_1 & \text{for} \quad y_i = 0, z_{1i} = 1 \\ \pi_2 & \text{for} \quad y_i = k, z_{2i} = 1 \\ \pi_3\, p_{yi}(\lambda_i) & \text{for} \quad y_i = 0, 1, \ldots,\ z_{3i} = 1. \end{cases}$$

where $p_{yi}(\lambda_i)$ is the Poisson probability mass function with mean λ_i.

Therefore the complete data likelihood function of the ZkIP model is

$$L_{comp}(\pi_1, \pi_2, \boldsymbol{\lambda}|\mathbf{y}, \mathbf{z}) = \prod_{i:y_i=0} (\pi_1)^{z_{1i}} \prod_{i:y_i=k} (\pi_2)^{z_{2i}} \prod_{i=1}^{n} (\pi_3 p_{yi}(\lambda_i))^{z_{3i}},$$

and the loglikelihood of the complete data, $(\mathbf{y}, \mathbf{z})$ for the ZkIP model is

$$\begin{aligned}
\ell_{comp}(\pi_1, \pi_2, \boldsymbol{\lambda}|\mathbf{y}, \mathbf{z}) &= \sum_{i:y_i=0} \big(z_{1i}\pi_1 + z_{3i}(\log \pi_3 + \log p_{0i}(\lambda_i))\big) \\
&+ \sum_{i:y_i=k} \big(z_{2i} \log \pi_2 + z_{3i}(\log \pi_3 + \log p_{ki}(\lambda_i))\big) \\
&+ \sum_{i=1}^{n} \big(z_{3i} \log \pi_3 + \log p_{yi}(\lambda_i)\big) \\
&= \sum_{i=1}^{n} (z_{1i}\gamma + z_{2i}\delta - \log(1 + e^{\gamma} + e^{\delta})) + \sum_{i=1}^{n} z_{3i} \log p_{yi}(\lambda_i).
\end{aligned} \tag{32}$$

When $\pi_2 = 0$, the ZkIP is reduced to the ZIP model. From (32), the loglikelihood of the ZIP for the complete data is

$$\begin{aligned}
\ell_{comp}(\pi_1, \boldsymbol{\lambda}|\mathbf{y}, \mathbf{z_1}) &= \sum_{i:y_i=0} \big(z_{1i}\pi_1 + (1 - z_{1i})(\log(1 - \pi_1) + \log p_{0i}(\lambda_i))\big) \\
&+ \sum_{i:y_i>0} \big((1 - z_{1i}) \log(1 - \pi_1) + \log p_{yi}(\lambda_i)\big) \\
&= \sum_{i=1}^{n} (z_{1i}\gamma - \log(1 + e^{\gamma}) + \sum_{i=1}^{n} (1 - z_{1i}) \log p_{yi}(\lambda_i).
\end{aligned} \tag{33}$$

Note that Lambert (1992) used equation (33) as the complete data loglikelihood for the ZIP model to get the EM estimates.

We now proceed to describe the EM algorithm for the ZkIP model. The first step in the EM algorithm involves selecting some starting values for the unknown parameters. The choice of the initial values are important for the convergence of the algorithm. A wrong choice of the initial values could result in slow convergence or breakdown of the algorithm. We recommend using the proportions of zeros and k's

respectively from the observed data as initial values for the parameters π_1 and π_2 and use the relations (29) to get initial values γ_0 and δ_0 for the parameters γ and δ respectively. The next step involves filling the latent values $\mathbf{z_i}$ by their expectations, which is the E-step. We will use the conditional expected values of $E(\mathbf{z}|\mathbf{y})$ given in Table 6 to generate $\mathbf{z_i}$'s. Recall that Table 6 is a reparametrized version of Table 1 given in Chapter 2.

Table 6. $E(\mathbf{z}|\mathbf{y})$ for the ZkIP regression model

$\mathbf{z}$	$y=0$	$y=k$	$y\neq 0,k$
z_1	$\frac{e^{\gamma}}{e^{\gamma}+p_{0i}(\lambda_i)}$	0	0
z_2	0	$\frac{e^{\delta}}{e^{\delta}+p_{ki}(\lambda_i)}$	0
z_3	$\frac{p_{0i}(\lambda_i)}{e^{\gamma}+p_{0i}(\lambda_i)}$	$\frac{p_{ki}(\lambda_i)}{e^{\delta}+p_{ki}(\lambda_i)}$	1

NOTE: The sum of entries in any column is one.

We use Table 6 to estimate the missing values in the expectation step of the EM algorithm as follows

$$\begin{aligned}
\widehat{z_{1i}} = E(z_{1i}|y_i=0) &= \frac{e^{\gamma}}{e^{\gamma}+p_{0i}(\lambda_i)} \quad \text{and} \quad \widehat{z_{1i}} = E(z_{1i}|y_i=k) = 0, \\
\widehat{z_{2i}} = E(z_{2i}|y_i=k) &= \frac{e^{\delta}}{e^{\delta}+p_{ki}(\lambda_i)} \quad \text{and} \quad \widehat{z_{2i}} = E(z_{21i}|y_i\neq k) = 0. \qquad (34)
\end{aligned}$$

For the maximization step in the EM algorithm, instead of maximizing the complete likelihood directly, we solve the score equations

$$\begin{aligned}
\frac{\partial \ell_{comp}}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \widehat{z_{3i}}(y_i - e^{\mathbf{x_i^T}\boldsymbol{\beta}})\,\mathbf{x_i} = 0 \\
\frac{\partial \ell_{comp}}{\partial \gamma} &= \sum_{i=1}^{n} \widehat{z_{1i}} - \frac{ne^{\gamma}}{(1+e^{\gamma}+e^{\delta})} = 0 \qquad (35) \\
\frac{\partial \ell_{comp}}{\partial \delta} &= \sum_{i=1}^{n} \widehat{z_{2i}} - \frac{ne^{\delta}}{(1+e^{\gamma}+e^{\delta})} = 0,
\end{aligned}$$

where $\widehat{z_{3i}} = (1 - \widehat{z_{1i}} - \widehat{z_{1i}})$. In summary the EM algorithm to estimate the parameters γ, δ and the regression parameter $\boldsymbol{\beta}$ for the ZkIP regression model is as follows.

1. Select initial values $\boldsymbol{\beta}_0$, γ_0, δ_0 for the parameters $\boldsymbol{\beta}$, γ, and δ respectively.

2. E-step: Estimate $\widehat{z_{1i}}$, $\widehat{z_{2i}}$ using equations (34).

3. M-step: Solve the score equations (35) and obtain an updated estimates $\boldsymbol{\beta}_1$, γ_1, δ_1.

4. Repeat the E-step and the M-step until the parameter estimates converge.

In the next section we will discuss how to obtain the standard errors of the estimates obtained by the EM algorithm.

3.4 STANDARD ERRORS FOR THE EM ESTIMATES

The most commonly used method to get the standard errors in the mixture models is to compute the matrix of partial derivatives of the loglikelihood for the observed data, that is, to calculate the information matrix from the observed data. Lambert (1992) used this method for computing the standard errors for ZIP regression model. Lin and Tsai (2012) used the Hessian matrix to get the standard errors for the ZkIP model without actually computing second order partial derivatives of the loglikelihood. Recall that the Hessian matrix comes out as a byproduct of the nonlinear optimization methods in the statistical software.

However, for the EM framework that we have used, an appropriate and easier approach for obtaining the standard errors is the method outlined by Louis (1982) that we had discussed in Section 2.5.2. The Louis (1982) method is based on the complete and missing data loglikeihoods and it is given by the relation (21) in Section 2.5.2 and for convenience we reproduce that equation here.

$$\widehat{\mathcal{I}_{obs}} = E(\mathcal{I}_{comp}|\mathbf{y}) - E(\mathcal{I}_{miss}|\mathbf{y}). \tag{36}$$

Recall that the loglikelihood of the complete data for the ZkIP regression model is given by (32) and the first order derivatives of this loglikelihood are given in equations (35). The elements of the matrix $E(\mathcal{I}_{comp}|\mathbf{y})$ are the expected values of the negative of second order partial derivatives of the complete data loglikelihood (32) and they are given by

$$\begin{aligned}
E\left[\frac{-\partial^2 \ell_{comp}}{\boldsymbol{\partial\beta\partial\beta}^T}\right] &= \sum_{i=1}^{n} \frac{[p_{0i}(\lambda_i)\, p_{ki}(\lambda_i) - e^{\gamma+\delta}]\, \lambda_i}{[e^{\gamma} + p_{0i}(\lambda_i)]\, [e^{\delta} + p_{ki}(\lambda_i)]} (\mathbf{x_i x_i^T}) \\
E\left[\frac{-\partial^2 \ell_{comp}}{\partial\gamma^2}\right] &= \frac{ne^{\gamma}(1+e^{\delta})}{(1+e^{\gamma}+e^{\delta})^2} \\
E\left[\frac{-\partial^2 \ell_{comp}}{\partial\gamma\partial\delta}\right] &= \frac{-ne^{\gamma+\delta}}{(1+e^{\gamma}+e^{\delta})^2} \\
E\left[\frac{-\partial^2 \ell_{comp}}{\partial\delta^2}\right] &= \frac{ne^{\delta}(1+e^{\gamma})}{(1+e^{\gamma}+e^{\delta})^2}.
\end{aligned}$$

The two elements $-\partial^2 \ell_{comp}/\boldsymbol{\partial\beta}\partial\gamma$ and $-\partial^2 \ell_{comp}/\boldsymbol{\partial\beta}\partial\delta$ are equal to zero and the other elements are obtained by symmetry. The loglikelihood of the missing data for the ZkIP regression model is

$$\begin{aligned}
\ell_{miss}(\boldsymbol{\beta},\, \gamma,\, \delta) &= \sum_{i=1}^{n} (z_{1i}\gamma + z_{2i}\delta + z_{3i} \log p_{yi}(\lambda_i)) - \sum_{i:y_i=0} \log(e^{\gamma} + p_{0i}(\lambda_i)) \\
&- \sum_{i:y_i=k} \log(e^{\delta} + p_{ki}(\lambda_i)) - \sum_{i:y_i\neq 0,k} \log p_{yi}(\lambda_i). \qquad (37)
\end{aligned}$$

The elements of the matrix $E(\mathcal{I}_{miss}|_{observed})$ are the negative of the expected value of second order derivatives of (37). These are given by the following equations

$$
\begin{aligned}
E\left[\frac{-\partial^2 \ell_{miss}}{\boldsymbol{\partial\beta\partial\beta}^T}\right] &= -E\Bigg[\Bigg(\sum_{i=1}^{n} z_{3i}\lambda_i \mathbf{x_i x_i^T} - \sum_{y_i\neq 0,k} \lambda_i \mathbf{x_i x_i^T} \\
&\quad - \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}(1-\lambda_i) + {p_{0i}}^2}{(e^{\gamma}+p_{0i})^2}\lambda_i \mathbf{x_i x_i^T} \\
&\quad - \sum_{i:y_i=k} \frac{e^{\delta} p_{ki}(\lambda_i - (k-\lambda_i)^2) + {p_{ki}}^2\lambda_i}{(e^{\delta}+p_{ki})^2}\mathbf{x_i x_i^T}\Bigg)\Big|\mathbf{y}\Bigg] \\
&= \sum_{i=1}^{n} \frac{p_{0i}p_{ki} - e^{\gamma+\delta}}{(e^{\gamma}+p_{0i})(e^{\delta}+p_{ki})}\lambda_i \mathbf{x_i x_i^T} - \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}(1-\lambda_i) + {p_{0i}}^2}{(e^{\gamma}+p_{0i})^2}\lambda_i \mathbf{x_i x_i^T} \\
&\quad - \sum_{i:y_i=k} \frac{e^{\delta} p_{ki}(\lambda_i - (k-\lambda_i)^2) + {p_{ki}}^2\lambda_i}{(e^{\delta}+p_{ki})^2}\mathbf{x_i x_i^T}
\end{aligned}
$$

and

$$
\begin{aligned}
E\left[\frac{-\partial^2 \ell_{miss}}{\boldsymbol{\partial\beta}\partial\gamma}\right] &= \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}\lambda_i \mathbf{x_i}}{(e^{\gamma}+p_{0i})^2} \\
E\left[\frac{-\partial^2 \ell_{miss}}{\boldsymbol{\partial\beta}\partial\delta}\right] &= -\sum_{i:y_i=k} \frac{e^{\delta} p_{ki}(k-\lambda_i)\mathbf{x_i}}{(e^{\delta}+p_{ki})^2} \\
E\left[\frac{-\partial^2 \ell_{miss}}{\partial\gamma^2}\right] &= \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}}{(e^{\gamma}+p_{0i})^2} \\
E\left[\frac{-\partial^2 \ell_{miss}}{\partial\gamma\partial\delta}\right] &= 0, \quad E\left[-\frac{\partial^2 \ell_{miss}}{\partial\delta^2}\right] = \sum_{i:y_i=k} \frac{e^{\delta} p_{ki}}{(e^{\delta}+p_{ki})^2}.
\end{aligned}
$$

All the above expected values are taken with respect to the missing values conditional on the observed data. Using these formulas we can compute $\widehat{\mathcal{I}_{obs}}$ given in equation (36). The square root of diagonal elements of $\left(\widehat{\mathcal{I}_{obs}}\right)^{-1}$ are the standard errors of the EM estimates.

3.5 EXAMPLES

In this section we illustrate the results presented in Sections 3.3 and 3.4 on two real life data. These data examples were obtained from the National Health Interview Survey (NHIS) conducted by the National Center for Health Sciences (NCHS). Since

1957, NCHS has been collecting and archiving data on US residents. The data is collected annually on various health topics including immunizations, depression, hepatitis, cancer, use of tobacco and other variables related to health. For our illustration we took a subset of data that was collected in year 2015. In both examples we fit zero and k inflated Poisson (ZkIP) model and compare it to the zero inflated Poisson (ZIP) and ordinary Poisson models. The first example illustrates a ZkIP model with inflations at 0 and $k = 6$, while the second example has inflations at 0 and $k = 1$, and the model is also known as zero and one inflated Poisson (ZOIP) model.

3.5.1 PAP SMEAR DATA

We revisit the pap smear data discussed in Section 2.7.1. In this example, we looked at the number of pap smear tests a female took in last six years for females aged more than 18 years. The data also consists of age of the female respondent and her answer to the question *ever received HPV shot or vaccine*?. The age is a continuous variable whereas the response to *HPV shot/vaccine* is a dichotomous variable. Both these variables could be treated as covariates in the model. As we mentioned before these data were obtained from NHIS adult survey files.

The mean number of the pap smear tests for this clean data is 3.40 and the variance is 5.25. The percentage (count) of females who never took a pap smear test was 15.68% (1884) and the percentage (count) of females who had one pap smear each year for a total of six in the last six years was 29.17% (3504). The observed frequencies are presented in Table 8. Clearly the proportions of zero and six in the data set are inflated and both these proportions are more than what we would expect under a Poisson model. Thus an appropriate model for this data is the zero and six inflated Poisson model or the ZkIP model with $k = 6$.

For the pap smear data using the methods described in Sections 3.3 and 3.4 we fit the ZkIP, ZIP and the Poisson models. We tested the significance of the two covariates in the models using Wald's test. It turned out that the variable age was not significant in the ZkIP and ZIP models. Age was removed in subsequent analysis and we reran the models with only *HPV shot* as the covariate. The shape of the loglikelihood function fixing the regression parameter for various values of π_2, π_1

for the ZkIP model is shown in Figure 5. As seen in the figure the loglikelihood is concave and appears to have a unique peak. The regression parameter is significant for all the models at $\alpha = 0.10$. The estimates obtained by the EM algorithm and the corresponding standard errors for the EM estimates described in Section 3.4 are presented in Table 7. For the ZkIP model the mixing parameter estimates were $\widehat{\pi}_1 = 0.126$, and $\widehat{\pi}_2 = 0.26$, meaning about 12.6% of the zeros were from the degenerate distribution and 26% of the observed frequencies of six pap smear count were from a degenerate distribution at six. The table also has the AIC value and maximum value of the loglikelihood function for different models. The AIC value of the ZkIP, ZIP and Poisson models are $46523.89, 52205.70, 56061.88$ respectively. The ZkIP model has minimum AIC and the difference between the AIC of ZkIP and ZIP model is greater than 5000. Thus, adding one more distribution which is degenerate at six to the model or the ZkIP with $k = 6$ is a better model than the ZIP for this data.

Table 7. Estimates and SE for pap smear

Parameter	ZkIP	ZIP	Poisson
Intercept	1.0837*	1.3696*	1.2192*
	(0.0086)	(0.0054)	(0.0053)
HPV shots	0.0727*	0.0333*	0.0276*
	(0.0235)	(0.0154)	(0.0152)
$\widehat{\gamma}$	-1.5844	-1.8132	–
	(0.0331)	(0.0184)	
$\widehat{\delta}$	-0.8526	–	–
	(0.0235)		
$\widehat{\pi}_1$	0.1257	0.1402	–
	(0.0026)	(0.0022)	
$\widehat{\pi}_2$	0.2613	–	–
	(0.0018)		
$\log L_{obs}$	-25363.93	-26098.85	-28028.94
AIC	46523.89	52205.70	56061.88

NOTE: The regression parameters significant at $\alpha = 0.10$ are asterisk marked. The standard errors are in parenthesis.

Recall that the three models Poisson, ZIP and ZkIP are nested models and we could use the likelihood ratio criterion described in the Section 2.6 to decide whether the complex model could be reduced to the simpler model. The LRT statistic which compares Poisson model with the ZIP is $-2\log\Lambda = 3860.18$ and the p-value computed using the limiting distribution, which is a mixture of two χ^2's with equal weights, is less than 0.0001. This implies that the inflation at zero is significant

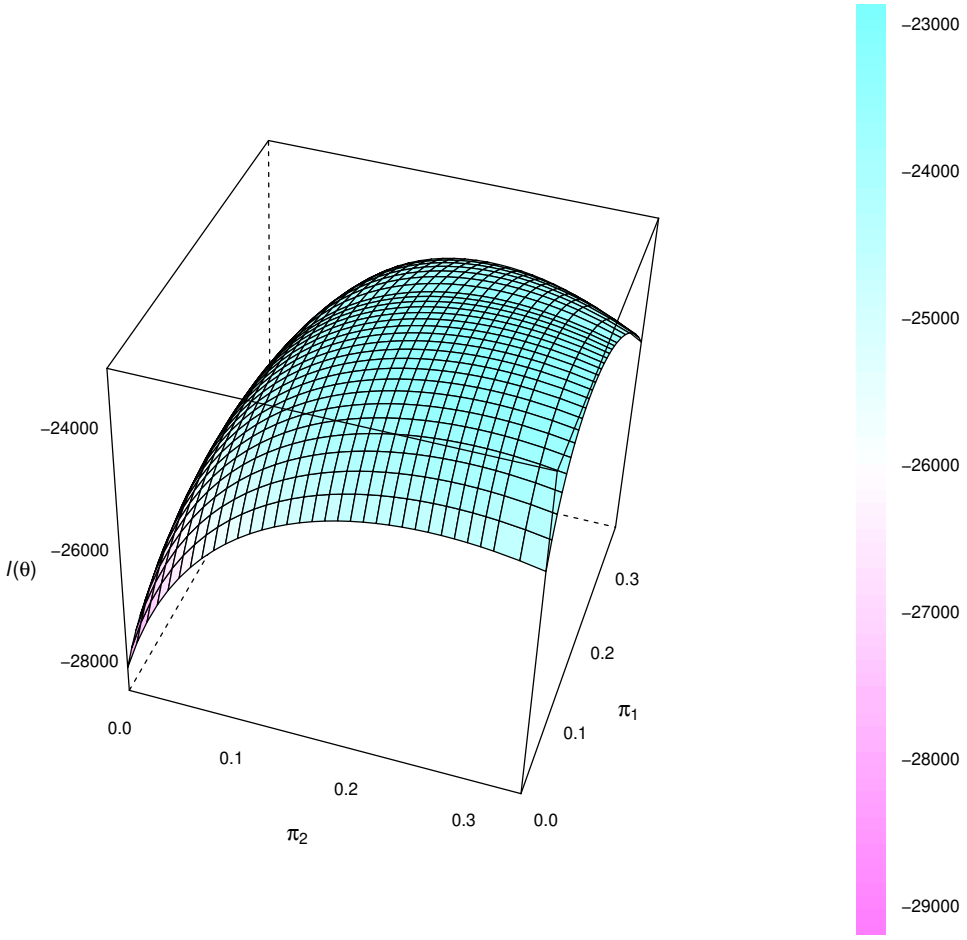


Figure 5. Loglikelihood of the ZkIP model for observed pap smear data.

and ZIP model is significantly better than the Poisson model. Similarly, we also used LRT to compare ZkIP with ZIP model. The value of the test statistic is $-2\log\Lambda = 1469.85$, which is again highly significant with a p-value less than 0.0001. Hence, ZkIP is significantly better than the ZIP model. Further, we checked the goodness-of-fit of the models by comparing the observed frequencies and the expected frequencies. Table 8 shows that the Poisson model has highest ABE and does not provide a good fit to the data. The error 5685.69 of the ZIP model is lower than that of the Poisson model 8086.16. And the sum of the absolute difference between the observed and expected frequency is minimum (1130.93) for the ZkIP model. Figure 6 shows that the ZkIP model is a good fit to the observed data. Thus the ZkIP which is able to capture inflated frequencies at both zero and 6 is a superior model for this data compared with ZIP and the Poisson model.

Table 8. Frequency comparisons for pap smear

Count	Observed	ZkIP	ZIP	Poisson
0	1884	1884.24	1883.47	402.81
1	1417	1112.73	785.54	1366.85
2	1362	1661.47	1553.79	2323.12
3	1536	1648.59	2043.78	2627.90
4	1115	1228.13	2016.95	2230.12
5	905	732.45	1592.78	1514.29
6	3504	3504.14	1048.87	857.41
> 6	291	162.46	600.28	421.80
ABE		1130.93	5685.69	8086.16
χ^2		297.64	7263.93	15312.36

3.5.2 EMERGENCY ROOM DATA

The data for this example was taken from the NHIS 2015 database on children aged less than 18 years. The count variable in this data is the number of visits of children to an emergency room (ER) in an year. For the covariates we chose age (0-17) and gender (Male/Female). We have removed the cases were the response or the covariates are missing, and ended up with a clean set of data for $n = 12223$ children. The average number of visits to the ER in our sample was 0.26, and the variance was 0.45. In the data the count values were 0 and 1 have frequencies 10046 and 1466. These frequencies are high because they account for 82.19 and 11.99 percentages of the total sample.

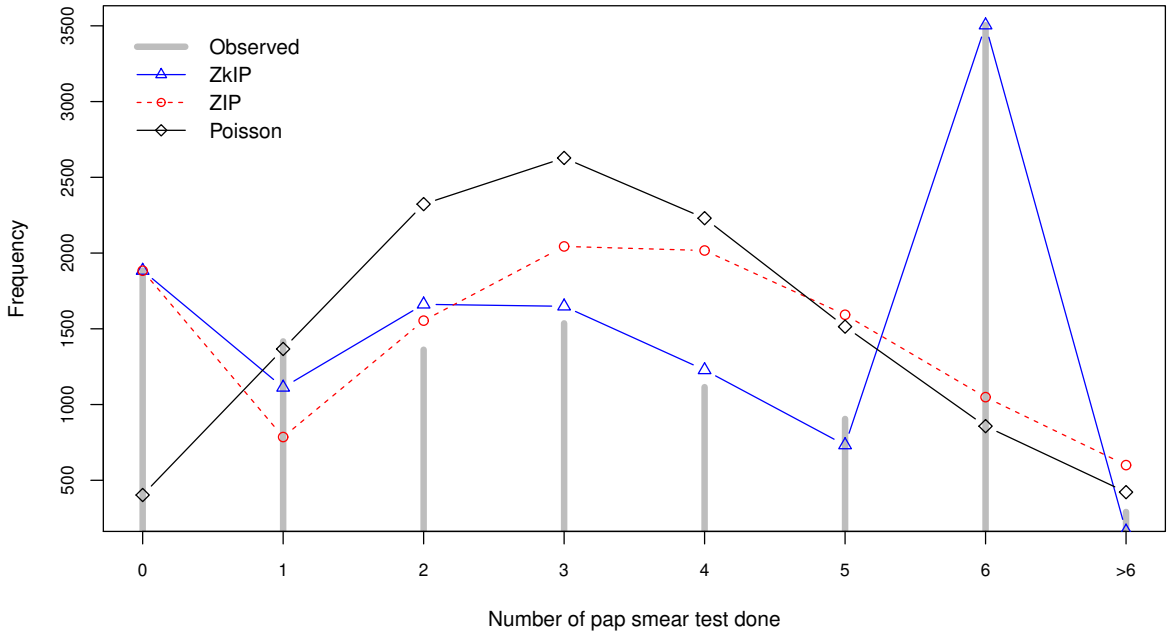


Figure 6. Observed and Expected Frequencies for pap smear data.

For this data we fit zero and one inflated Poisson (ZOIP), zero inflated Poisson (ZIP), and the Poisson model. The significance of the regression variables is tested using Wald Test. In the first iteration, the gender variable was found to be insignificant in all three models so it was removed from the models. The analysis is again performed with only age as the covariate. The model estimates and standard errors are presented in Table 9.

Table 9. Estimates and SE for ER data

Paramter	ZkIP	ZIP	Poisson
Intercept	0.1173*	-0.0314*	-1.0512*
	(0.0610)	(0.0395)	(0.0312)
Age	-0.0217*	-0.0252*	-0.0358*
	(0.0044)	(0.0039)	(0.0033)
$\widehat{\gamma}$	1.0959	0.7098	–
	(0.1210)	(0.0427)	
$\widehat{\delta}$	-2.0450	–	–
	(0.3853)		
$\widehat{\pi}_1$	0.7260	0.6704	–
	(0.0213)	(0.0094)	
$\widehat{\pi}_2$	0.0314	–	–
	(0.0679)		
$\log L_{obs}$	-7736.62	-7741.19	-8295.23
AIC	15481.24	15488.39	16594.00

NOTE: The regression parameters significant at $\alpha = 0.05$ are asterisk marked. The standard errors are given in parenthesis.

The AIC value of the ZOIP, ZIP and Poisson models are $15481.24, 15488.39, 16594.00$ respectively. Using the AIC measure, we see the ZIP model performs better than the Poisson model. Further, the ZOIP model seems to have a slight edge over the ZIP model. We also performed the likelihood ratio test for model selection. The LRT statistic for testing Poisson model over ZIP is given by $-2\log\Lambda = 1108.08$, which is highly significant. The LRT statistic $-2\log\Lambda = 9.15$ shows that the ZOIP model is significantly better than the ZIP. Thus both the AIC and LRT criterion shows that ZOIP fits best for this data.

The observed and expected frequencies of the ZOIP, ZIP and Poisson models are in Table 10 and they are plotted in Figure 7. The ZIP model is able to capture the inflation at count zero. However, the ZOIP model is able to capture the inflation at count zero and one as well. The conclusion is also supported by the ABE measure.

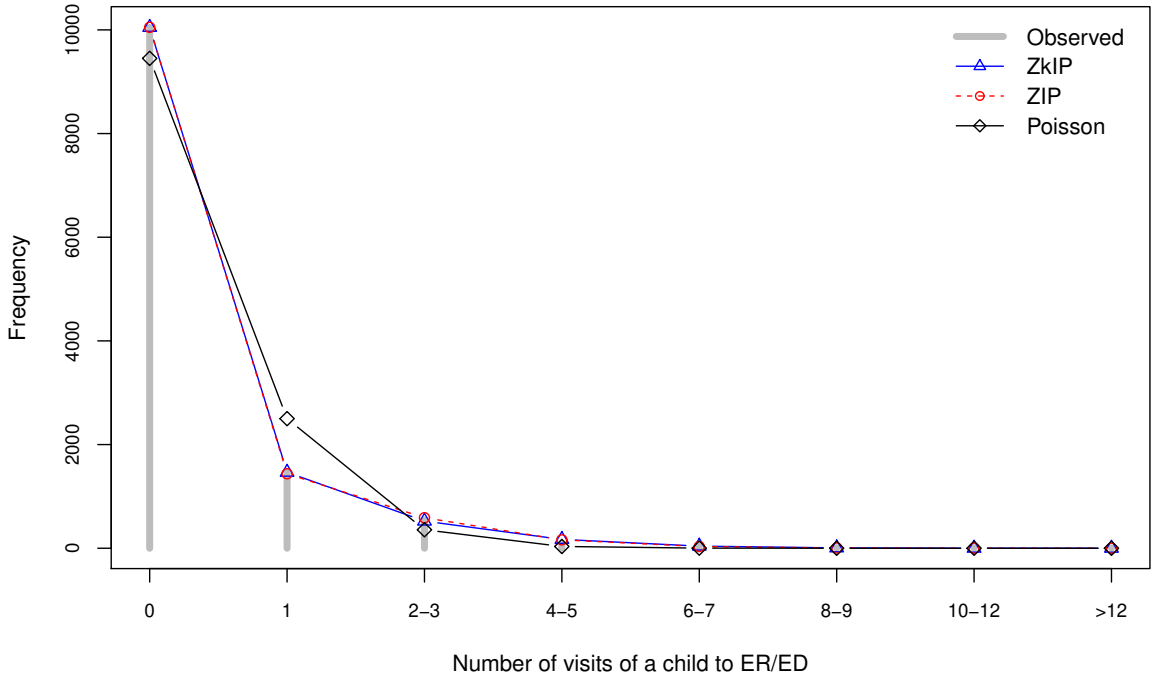


Figure 7. Observed and Expected Frequencies for ER data.

Table 10. Frequency comparisons for ER data

Count	Observed	ZOIP	ZIP	Poisson
0	10046	10047.60	10049.72	9450.22
1	1466	1465.98	1436.96	2499.80
2-3	548	523.82	588.65	356.52
4-5	92	170.26	162.21	34.31
6-7	37	41.01	33.08	2.43
8-9	12	6.92	4.65	0.12
10-12	13	0.86	0.47	0.00
> 12	9	0.21	0.10	0.00
ABE		134.07	176.32	1947.20
χ^2		586.40	1150.84	304381.20

CHAPTER 4

ZERO AND K INFLATED CONWAY-MAXWELL-POISSON MODELS

4.1 INTRODUCTION

In Chapter 3 we constructed ZkIP regression model for a two point inflated count data. The ZkIP is a mixture of three distribution, one of which is the Poisson. In this chapter, we replace the Poisson with Conway-Maxwell-Poisson (CMP), which is a two parameter extension of the ordinary Poisson. Before embarking on this generalization we first introduce the CMP distribution and its basic properties. A nice summary of the CMP distribution can also be found in Shmueli et al. (2005). In a recent paper, Sellers and Raim (2016) introduced the zero inflated CMP (ZICMP), which extends the CMP to handle excess zeros in the data. We give a brief summary of ZICMP model. While this model can handle excess zeros it is not an appropriate model when there is another count with a high frequency. To handle inflations at two points zero and k, we introduce in this chapter zero and k inflated Conway-Maxwell-Poisson (ZkICMP) distribution. This ZkICMP distribution extends ZkIP and it is much more flexible model to account two point inflations in the count data. We study first basic properties of ZkICMP including a stochastic representation. To study the relationship between the count response and explanatory variables we construct the ZkICMP regression model. We discuss estimation of the regression parameter and the mixing probabilities using maximum likelihood variable in Section 4.3.1. The formulas needed to calculate the standard errors are given in Section 4.3.2. In the following section we discuss inferential issues and model selection. We also study performance of the ZkICMP using two simulated data sets. We conclude the chapter with application of the ZkICMP regression model on two real-life examples from National Health and Nutrition Examination Survey (NHANES).

4.2 ZKICMP PROBABILITY DISTRIBUTION

We say that a count response Y follows the Conway-Maxwell-Poisson (CMP) distribution if its probability mass function is given by

$$P(Y = y) = \frac{\lambda^y}{(y!)^\nu \, Z(\lambda, \nu)}, \quad \text{for } y = 0, 1, 2, \ldots \tag{38}$$

where

$$Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu} \quad \text{and } \nu \geq 0, \ \ \lambda > 0.$$

Here, λ is the rate parameter and ν is the dispersion parameter. Note, the function $Z(\lambda, \nu)$ is an infinite series. We can check that this series converges for any $\lambda > 0, \nu > 0$ and when $\nu = 0$ it converges for $0 < \lambda < 1$. Recently, Shmueli et al. (2005) brought the CMP distribution into the limelight. They gave various statistical properties of the distribution, and we will discuss those properties here briefly. The dispersion parameter $\nu = 1$ corresponds to equidispersion and the CMP distribution reduces to the Poisson distribution. While $\nu < 1$ corresponds to overdispersion and $\nu > 1$ indicates underdispersion. Further, when $\nu = 0$ and $\lambda < 1$ the CMP becomes the geometric distribution with success probability $p = (1 - \lambda)$. And the CMP distribution converges to Bernoulli$(p = \lambda/(1 + \lambda))$ distribution when $\nu \to \infty$. The moment generating function of the distribution (38) is given by

$$M_Y(t) = \frac{Z(\lambda e^t, \nu)}{Z(\lambda, \nu)}.$$

It could be used to derive the raw moments of the distribution. The mean and variance of the CMP distribution are

$$E(Y) = \lambda \frac{\partial \log Z}{\partial \lambda}, \qquad V(Y) = \frac{\partial E(Y)}{\partial \log \lambda}.$$

Shmueli et al. (2005) gave the following approximations

$$\begin{aligned} E(Y) &\approx \lambda^{1/\nu} - \frac{\nu - 1}{2\nu} \\ V(Y) &\approx \frac{1}{\nu} \lambda^{1/\nu}. \end{aligned}$$

These approximations are good when $\nu \leq 1$ or $\lambda > 10^{\nu}$. Another useful function is the probability generating function. For the CMP this function is

$$G_Y(t) = Z(\lambda t, \nu)/Z(\lambda, \nu).$$

The CMP distribution belongs to the exponential family. Indeed we can rewrite (38) as

$$P(Y = y) \;=\; \exp\left(y \log \lambda - \nu \log y! - \log Z(\lambda, \nu)\right).$$

Clearly, the sufficient statistics are $(y, \log y!)$ for (λ, ν). The CMP distribution also belongs to the scale family. If $X = \sigma Y$ then

$$\begin{aligned} P(X = y) &= \frac{1}{\sigma} \frac{\lambda^{y/\sigma}}{(y/\sigma!)^{\nu} Z(\lambda, \nu)} \\ &= \frac{1}{\sigma}\, CMP\left(\frac{y}{\sigma}\right). \end{aligned}$$

An appropriate model for underdispersed or overdispersed count data with excessive zeros is the zero inflated Conway-Maxwell-Poisson (ZICMP) distribution that was introduced by Sellers and Raim (2016). The ZICMP distribution is an extension of the ZIP distribution that we discussed in Chapter 2. It is a mixture of degenerate distribution at zero with probability π_1 and a CMP distribution with probability $(1 - \pi_1)$. The probability mass function of ZICMP distribution is

$$P(Y = y) = \begin{cases} \pi_1 + (1 - \pi_1)\, \dfrac{1}{Z(\lambda, \nu)} & \text{when } y = 0 \\ (1 - \pi_1)\, \dfrac{\lambda^y}{(y!)^{\nu} Z(\lambda, \nu)} & \text{when } y \geq 1. \end{cases}$$

where $0 < \pi_1 < 1$, $\lambda > 0$ and $\nu \geq 0$. The mean and variance of ZICMP distribution are

$$\begin{aligned} E(Y) &= (1 - \pi_1)\, \lambda\, \frac{\partial \log Z(\lambda, \nu)}{\partial \lambda}, \\ V(Y) &= (1 - \pi_1) \left(\lambda\, \frac{\partial E(Y)}{\partial \lambda} + \pi_1\, (E(Y))^2\right). \end{aligned}$$

An extension of ZICMP is the ZkICMP distribution. Similar to ZkIP, it is a mixture of three distributions including point masses at 0 and k. The probability mass function of ZkICMP distribution is

$$P(Y=y)=\begin{cases}\pi_1+(1-\pi_1-\pi_2)\,\dfrac{1}{Z(\lambda,\nu)} & \text{when } y=0\\ \pi_2+(1-\pi_1-\pi_2)\,\dfrac{\lambda^k}{(k!)^\nu Z(\lambda,\nu)} & \text{when } y=k\\ (1-\pi_1-\pi_2)\,\dfrac{\lambda^y}{(y!)^\nu Z(\lambda_i,\nu)} & \text{when } y\geq 1,\ y\neq k,\end{cases} \tag{39}$$

where $0<\pi_1<\pi_1+\pi_2<1$, $\lambda>0$ and $\nu\geq 0$. The mean and variance of (39) are

$$\begin{aligned}E(Y) &= k\pi_2+\pi_3\lambda\frac{\partial\log Z}{\partial\lambda}\\ V(Y) &= \pi_2\left(k^2(1-\pi_2)-2k\pi_3\lambda\frac{\partial\log Z(\lambda,\nu)}{\partial\lambda}\right)\\ &\quad +\pi_3\left(\frac{\partial E(Y)}{\partial\log\lambda}+(1+\pi_3)(\lambda\frac{\partial\log Z(\lambda,\nu)}{\partial\lambda})^2\right).\end{aligned}$$

where $\pi_3=(1-\pi_1-\pi_2)$. The moment generating function is

$$M_Y(t)=\pi_1+\pi_2e^{tk}+\pi_3\,\frac{Z(\lambda e^t,\nu)}{Z(\lambda,\nu)},$$

and the probability generating function is

$$G_Y(t)=\pi_1+\pi_2t^k+\pi_3\,\frac{Z(\lambda t,\nu)}{Z(\lambda,\nu)}.$$

When $\nu=1$, we have $Z(\lambda,\nu)=e^\lambda$ and (39) reduces to

$$P(Y=y)=\begin{cases}\pi_1+\pi_3e^{-\lambda} & \text{when } y=0\\ \pi_2+\pi_3\dfrac{\lambda^ke^{-\lambda}}{k!} & \text{when } y=k\\ \pi_3\dfrac{\lambda^ye^{-\lambda}}{y!} & \text{when } y\geq 1, y\neq k,\end{cases}$$

which is the ZkIP that we discussed in Chapter 2. When $\nu=0$, we have $Z(\lambda,\nu)=$

$1/(1-\lambda)$ and ZkICMP reduces

$$P(Y=y) = \begin{cases} \pi_1 + \pi_3(1-\lambda) & \text{when } y = 0 \\ \pi_2 + \pi_3\lambda^k(1-\lambda) & \text{when } y = k \\ \pi_3\lambda^y(1-\lambda) & \text{when } y \geq 1, y \neq k. \end{cases}$$

which is same as ZkIG. Next, $Z(\lambda,\nu) \to (1+\lambda)$ as $\nu \to \infty$, and therefore we have

$$\begin{aligned} P(Y=0) &\to \pi_1 + \pi_3 \frac{1}{1+\lambda}, \\ P(Y=1) &\to \pi_2 + \pi_3 \frac{\lambda}{1+\lambda}. \\ P(Y>1) &\to 0. \end{aligned}$$

Thus the ZkICMP distribution reduces to Bernoulli distribution with success probability $\lambda/(1+\lambda)$ as $\nu \to \infty$.

4.3 ZKICMP REGRESSION MODEL

In this section we will study the ZkICMP regression model. Assume that we have a vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ of n independent observations. Associated with each y_i we have a p dimensional covariate vector $\mathbf{x_i} = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$. We set $x_{i1} = 1$ for the regression model with an intercept. The general layout of the observed data is in Table 11. Assuming a possible model for y_i is the ZkICMP distribution with parameters $(\pi_1, \pi_2, \lambda_i, \nu)$, the likelihood function is

$$\begin{aligned} L_{obs}(\pi_1,\pi_2,\boldsymbol{\lambda},\nu|\mathbf{y}) &\propto \prod_{i:y_i=0}\left(\pi_1 + \pi_3\frac{1}{Z(\lambda_i,\nu)}\right)\prod_{i:y_i=k}\left(\pi_2 + \pi_3\,\frac{{\lambda_i}^k}{(k!)^\nu Z(\lambda_i,\nu)}\right) \\ &\quad \prod_{i:y_i\neq 0,k}\left(\pi_3\,\frac{{\lambda_i}^{y_i}}{(y_i!)^\nu Z(\lambda_i,\nu)}\right) \\ &\propto \prod_{i:y_i=0}(\pi_1 + \pi_3\,p_{0i})\prod_{i:y_i=k}(\pi_2 + \pi_3\,p_{ki})\prod_{i:y_i\neq 0,k}(\pi_3\,p_{y_i}), \end{aligned} \tag{40}$$

where $\pi_3 = (1-\pi_1-\pi_2)$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$ and $p_{y_i} = {\lambda_i}^{y_i}/[(y_i!)^\nu Z(\lambda_i,\nu)]$.

For the regression we set $\log(\lambda_i) = \mathbf{x_i^T}\boldsymbol{\beta}$, a linear function of the covariates and

Table 11. General Layout of the count data

Observations	Response	Covariates			
1	y_1	x_{11}	...	...	x_{1p}
2	y_2	x_{21}	...	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$			$\vdots$
i	y_i	x_{i1}	...	...	x_{ip}
$\vdots$	$\vdots$	$\vdots$			$\vdots$
n	y_n	x_{n1}	...	...	x_{np}

the regression parameter $\boldsymbol{\beta}$. While it is possible to link the parameters π_1, π_2 and ν to the covariates but for simplicity we assume they are unknown constants. For obtaining maximum likelihood estimates and to use the optimization routines we reparametrize them as follows

$$\log\left(\frac{\pi_1}{\pi_3}\right) = \gamma, \quad \log\left(\frac{\pi_2}{\pi_3}\right) = \delta, \quad \text{and} \quad \log(\nu) = \eta. \tag{41}$$

When $\pi_2 = 0$ the likelihood function (40) simplifies to

$$L_{obs}(\pi_1, \boldsymbol{\lambda}, \nu|\mathbf{y}) \propto \prod_{i:y_i=0} \left(\pi_1 + (1-\pi_1)\frac{1}{Z(\lambda_i,\nu)}\right) \prod_{i:y_i>0} \left((1-\pi_1)\,\frac{{\lambda_i}^{y_i}}{(y_i!)^{\nu} Z(\lambda_i,\nu)}\right),$$

which is the likelihood function of the ZICMP model (Sellers and Raim, 2016). When $\pi_2 = 0$ and $\pi_1 = 0$ equation (40) becomes

$$L_{obs}(\boldsymbol{\lambda}, \nu|\mathbf{y}) = \prod_{i=1}^{n} \frac{{\lambda_i}^{y_i}}{(y_i!)^{\nu} Z(\lambda_i,\nu)},$$

which is the likelihood function of the CMP distribution (Shmueli et al., 2005).

4.3.1 ESTIMATION OF ZKICMP PARAMETERS

In this section we discuss estimation of the ZkICMP regression model parameters given by the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}, \gamma, \delta, \nu)$. The dimension p of the regression parameter vector $\boldsymbol{\beta}$ depends on the number of covariates included in the model. The parameters

γ and δ quantify the zero and k inflations respectively. The dispersion parameter in the ZkICMP model is ν. Please note that both ZkICMP and ZkINB have the same number of parameters, and they have one less than ZICMP and ZkIP models.

The loglikelihood function of the observed data for the ZkICMP model is given by

$$\ell_{obs}(\boldsymbol{\theta}) = \sum_{i:y_i=0} \log(\pi_1 + \pi_3\, p_{0i}) + \sum_{i:y_i=k} \log(\pi_2 + \pi_3\, p_{ki}) + \sum_{i:y_i\neq 0,k} (\log \pi_3 + \log p_{y_i})$$

where $p_{y_i} = {\lambda_i}^{y_i}/[(y_i!)^{\nu} Z(\lambda_i, \nu)]$. Substituting $\pi_1 = e^{\gamma}/(1 + e^{\gamma} + e^{\delta})$ and $\pi_2 = e^{\delta}/(1 + e^{\gamma} + e^{\delta})$, the loglikelihood can be rewritten as

$$\begin{aligned}\ell_{obs}(\boldsymbol{\theta}) &= \sum_{i:y_i=0} \log(e^{\gamma} + p_{0i}) + \sum_{i:y_i=k} \log(e^{\delta} + p_{ki}) + \sum_{i:y_i\neq 0,k} \log p_{y_i} \\ &\quad -n(\log(1 + e^{\gamma} + e^{\delta})). \end{aligned} \tag{42}$$

Taking the partial derivatives of (42) with respect to the parameters we obtain the following score functions. For notational convenience we write Z instead of $Z(\lambda_i, \nu)$.

$$\begin{aligned}\frac{\partial \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \left(-\sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma} + p_{0i}} \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i} + \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta} + p_{ki}} \left(\frac{k}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i} \right) \right) \lambda_i \mathbf{x_i} \\ &\quad + \sum_{i:y_i\neq 0,k} \left(\frac{y_i}{\lambda_i} - \frac{1}{Z}\frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \\ \frac{\partial \ell_{obs}(\boldsymbol{\theta})}{\partial \gamma} &= \sum_{i:y_i=0} \frac{e^{\gamma}}{e^{\gamma} + p_{0i}} - \frac{n e^{\gamma}}{1 + e^{\gamma} + e^{\delta}} \end{aligned}$$

$$\begin{aligned}\frac{\partial \ell_{obs}(\boldsymbol{\theta})}{\partial \delta} &= \sum_{i:y_i=k} \frac{e^{\delta}}{e^{\delta} + p_{ki}} - \frac{n e^{\delta}}{1 + e^{\gamma} + e^{\delta}} \\ \frac{\partial \ell_{obs}(\boldsymbol{\theta})}{\partial \nu} &= -\sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma} + p_{0i}} \frac{1}{Z}\frac{\partial Z}{\partial \nu} - \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta} + p_{ki}} \left(\log(k!) + \frac{1}{Z}\frac{\partial Z}{\partial \nu} \right) \\ &\quad - \sum_{i:y_i\neq 0,k} \left(\log(y_i!) + \frac{1}{Z}\frac{\partial Z}{\partial \nu} \right) \end{aligned}$$

The score equations do not have closed form solutions, and thus have to be solved

numerically using routines for solving non-linear equations. There are several routines available in R software including *nlminb*, *optim*, box-constraint quasi-newton (L-BGFS-B), conjugate gradient (CG), simulated annealing algorithm (SANN). We have used *nlminb*, which essentially is an implementation of Newton-Raphson. The estimate of $\boldsymbol{\beta}$ that we obtain from fitting a Poisson model is taken as the initial value of $\boldsymbol{\beta}$. The initial values for γ, δ are obtained from the sample proportions of zeros and ks and we take one as the initial value for ν. We did not encounter any convergence problems with these initial values. Details for obtaining the standard errors of parameter estimates will be discussed in the next section.

4.3.2 STANDARD ERRORS FOR THE PARAMETER ESTIMATES

Under standard regularity conditions according to Cramèr's theorem, the maximum likelihood estimates are asymptotically normal with covariance matrix given by the inverse of Fisher information. For the ZkICMP model the Fisher information is given by

$$\mathcal{I}_{obs} = \left[-\frac{\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \right]$$

where $\ell_{obs}(\boldsymbol{\theta})$ is given by (42). The elements of $\mathcal{I}_{obs}$ can be obtained taking the partial derivatives of the score functions and they are given by

$$\begin{aligned}
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta^T}} &= -\sum_{i:y_i=0} \frac{p_{0i}}{e^\gamma + p_{0i}} \frac{1}{Z} \left(\frac{2\lambda_i}{Z} \left(\frac{\partial Z}{\partial \lambda_i} \right)^2 - \frac{p_{0i}}{e^\gamma + p_{0i}} \frac{\lambda_i}{Z} \frac{\partial Z}{\partial Z} - \frac{\partial \lambda_i}{\partial \lambda_i} - \lambda_i \frac{\partial^2 Z}{\partial {\lambda_i}^2} \right) \\
&\qquad \times \quad \lambda_i \mathbf{x_i} \mathbf{x_i^T} \\
&\quad - \sum_{i:y_i=k} \frac{p_{ki}}{e^\delta + p_{ki}} \left(\frac{e^\delta}{e^\delta + p_{ki}} \lambda_i \left(\frac{k}{\lambda_i} - \frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right)^2 \right) \lambda_i \mathbf{x_i} \mathbf{x_i^T} \\
&\quad - \sum_{i:y_i=k} \frac{p_{ki}}{e^\delta + p_{ki}} \left(\lambda_i \left(\left(\frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right)^2 - \frac{1}{Z} \frac{\partial^2 Z}{\partial {\lambda_i}^2} \right) - \frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \mathbf{x_i^T} \\
&\quad + \sum_{i:y_i \neq 0,k} \left(\lambda_i \left(\left(\frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right)^2 - \frac{1}{Z} \frac{\partial^2 Z}{\partial {\lambda_i}^2} \right) - \frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \mathbf{x_i^T}
\end{aligned}$$

$$
\begin{aligned}
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \gamma} &= -\sum_{i:y_i=0} \frac{e^{\gamma}}{(e^{\gamma}+p_{0i})^2} p_{0i} \lambda_i \mathbf{x_i} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \delta} &= \sum_{i:y_i=k} \frac{e^{\delta}}{(e^{\delta}+p_{ki})^2} p_{ki} (k-\lambda_i) \mathbf{x_i} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \nu} &= \sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma}+p_{0i}} \frac{1}{Z} \left(\frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z} \frac{\partial Z}{\partial \nu} \frac{\partial Z}{\partial \lambda_i} - \frac{e^{\gamma}}{e^{\gamma}+p_{0i}} \frac{1}{Z} \frac{\partial Z}{\partial \nu} \frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \\
&\quad + \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}} \left(\frac{1}{Z} \frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z} \frac{\partial Z}{\partial \nu} \frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \\
&\quad + \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}} \left(\frac{e^{\delta}}{e^{\delta}+p_{ki}} \left(\log k! + \frac{1}{Z} \frac{\partial Z}{\partial \nu} \right) \left(\frac{k}{\lambda_i} - \frac{1}{Z} \frac{\partial Z}{\partial \lambda_i} \right) \right) \lambda_i \mathbf{x_i} \\
&\quad + \sum_{i:y_i \neq 0,k} \frac{1}{Z} \left(\frac{\partial^2 Z}{\partial \nu \partial \lambda_i} - \frac{1}{Z} \frac{\partial Z}{\partial \nu} \frac{\partial Z}{\partial \lambda_i} \right) \lambda_i \mathbf{x_i} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \gamma^2} &= \frac{n e^{\gamma}(1+e^{\delta})}{(1+e^{\gamma}+e^{\delta})^2} - \sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}}{(e^{\gamma}+p_{0i})^2} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \gamma \partial \delta} &= \frac{-n e^{\gamma+\delta}}{(1+e^{\gamma}+e^{\delta})^2} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \gamma \partial \nu} &= -\sum_{i:y_i=0} \frac{e^{\gamma} p_{0i}}{(e^{\gamma}+p_{0i})^2} \frac{1}{Z} \frac{\partial Z}{\partial \nu}
\end{aligned}
$$

$$
\begin{aligned}
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \delta^2} &= \frac{n e^{\delta}(1+e^{\gamma})}{(1+e^{\gamma}+e^{\delta})^2} - \sum_{i:y_i=k} \frac{e^{\delta} p_{ki}}{(e^{\delta}+p_{ki})^2} \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \delta \partial \nu} &= -\sum_{i:y_i=k} \frac{e^{\delta} p_{ki}}{(e^{\delta}+p_{ki})^2} \left(\log k! + \frac{1}{Z} \frac{\partial Z}{\partial \nu} \right) \\
\frac{-\partial^2 \ell_{obs}(\boldsymbol{\theta})}{\partial \nu^2} &= \sum_{i:y_i=0} \frac{p_{0i}}{e^{\gamma}+p_{0i}} \frac{1}{Z} \left(\frac{\partial^2 Z}{\partial \nu^2} - \frac{p_{0i}+2e^{\gamma}}{e^{\gamma}+p_{0i}} \frac{1}{Z} \left(\frac{\partial Z}{\partial \nu} \right)^2 \right) \\
&\quad + \sum_{i:y_i=k} \frac{p_{ki}}{e^{\delta}+p_{ki}} \left(\frac{1}{Z} \frac{\partial^2 Z}{\partial \nu^2} - \left(\frac{1}{Z} \frac{\partial Z}{\partial \nu} \right)^2 - \frac{e^{\delta}}{e^{\delta}+p_{ki}} \left(\log k! + \frac{1}{Z} \frac{\partial Z}{\partial \nu} \right)^2 \right) \\
&\quad + \sum_{i:y_i \neq 0,k} \left(\frac{1}{Z} \frac{\partial^2 Z}{\partial \nu^2} - \left(\frac{1}{Z} \frac{\partial Z}{\partial \nu} \right)^2 \right)
\end{aligned}
$$

4.4 HYPOTHESIS TESTING AND MODEL SELECTION

In the previous section we have studied maximum likelihood estimation of the ZkICMP regression model parameters. Under standard regularity conditions, the parameter estimates are asymptotically normal with standard errors given by the inverse Fisher information. We can use this asymptotic result to construct test of hypothesis for the significance of the regression coefficients. For model selection as in Chapter 2, we can use Akaike Information Criterion (AIC) or the likelihood-ratio test statistic to select the model that best fits the data. The adequacy of the selected model can be validated using the goodness of fit criteria or by studying the residuals. We discuss testing of hypothesis in Section 4.4.1, residual analysis in Section 4.4.2, and model selection in Section 4.4.3.

4.4.1 HYPOTHESIS TESTING

Testing the impact of jth covariate on the count response is equivalent to testing $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$. This is straightforward and can be done using standard Wald's statistic, $z = \widehat{\beta}_j / SE(\widehat{\beta}_j)$, which is asymptotically standard normal under the null hypothesis. The other parameter of interest is the dispersion parameter ν. We would be interested in testing either $H_0 : \nu = 1$ against $H_1 : \nu > 1$ or $H_1 : \nu < 1$. The former hypothesis is indicative of overdispersion and the latter indicative of underdispersion in the data. Recall that when $\nu = 1$ the ZkICMP distribution is same as the ZkIP distribution. Therefore to check if we could replace ZkICMP model with ZkIP, we could test null hypothesis $H_0 : \nu = 1$ versus the two sided alternative $H_1 : \nu \neq 1$. Once again all these hypothesis could be tested using the Wald's test statistic, $z = \widehat{\nu} / SE(\widehat{\nu})$ which is asymptotically standard normal. An alternative for testing $H_0 : \nu = 1$ is the likelihood ratio test. The test statistic is

$$-2 \log \Lambda = -2 \log \frac{L_{obs}(\hat{\hat{\beta}},\, \hat{\hat{\gamma}},\, \hat{\hat{\delta}},\, \nu = 1)}{L_{obs}(\widehat{\beta}, \widehat{\gamma}, \widehat{\delta}, \widehat{\nu})}, \tag{43}$$

which is asymptotically distributed as chi-square with one degree of freedom. In the above $\hat{\hat{\beta}}$, $\hat{\hat{\gamma}}$ and $\hat{\hat{\delta}}$ are the ML estimates for the ZkIP model.

The LRT test could also be used to test for inflations at zero and at k. As in

Chapter 2, the asymptotic distribution of the LRT statistic is a mixture of chi-squares under the null hypothesis $H_0 : \pi_1 = 0$ or under the null hypothesis $H_0 : \pi_2 = 0$.

4.4.2 RESIDUAL ANALYSIS

For grouped data a popular measure of goodness-of-fit is the Pearson statistic. However, the measure is not suitable for ungrouped or subject specific data. The differences between the observed and fitted values for each subject or observation are best studied in the residual analysis. Dunn and Smyth (1996) showed that usual residuals may not be the best choice for non-normal models. And they recommend the use of randomized quantile residuals for non-normal responses, in particular, for count responses. The randomized quantile residuals for the ZICMP models were used by Sellers and Raim (2016) to check for goodness-of-fit. They developed 'COMPoissonReg' package in R to calculate the residuals for the ZICMP and CMP models. We have modified their code to calculate the quantile residuals for ZkICMP to check for goodness-of-fit in the examples discussed in Section 4.6.

4.4.3 MODEL SELECTION

We have seen there are various models ZkICMP, ZICMP, ZkINB, ZkIP, and ZIP, that one could use for a given count data with inflated frequencies. Besides test of hypothesis using the likelihood ratio test, we could select the best and appropriate model using the Akaike Information Criterion (AIC) defined in Section 2.6.4. The AIC safeguards over fitting by adding a penalty term for the number of parameters in the model. The best model is the one with minimum AIC value.

4.5 SIMULATIONS

In this section we check our parameter estimation methods for ZkICMP model on simulated count data that contains inflated frequencies at zero and at k. The data is simulated from ZkICMP distribution for various values of $\boldsymbol{\theta} = (\lambda, \nu, \pi_1, \pi_2)$. The value of $\nu > 1$ in all the simulations which symbolizes underdispersion. In the simulated data zeros and k's are inflated because π_1 and π_2 are positive. We tried various sample sizes $n = 200, 500, 1000, 2000$. We fit other count models for

the simulated data and show using various criterion ZkICMP fits best among the competing models.

4.5.1 SIMULATED DATA I

For our first simulated data sets we chose $\lambda = 8$, $\nu = 2$, $\pi_1 = 0.5$, and $\pi_2 = 0.2$. Thus fifty percent of the simulated data comes from the degenerate distribution at zero and twenty percent comes from the degenerate distribution at $k = 1$, and the rest from an underdispersed CMP distribution since $\nu = 2$. Four different sample sizes $n = 200$, 500, 1000, 2000 were used. We obtain the maximum likelihood estimates and standard errors of the parameters for various count models. The results are shown in Table 12 for various sample sizes n.

We observe for simulated data with sample size $n = 500$, the Poisson model gives an estimate $\widehat{\lambda} = 1.11$, which is far from the true value $\lambda = 8$. The ZIP model accounts for the inflation at zero but gives a poor estimate (1.80) for λ. The ZkIP and ZkINB models underestimate inflations both at zero ($\widehat{\pi}_1 = 0.45$) and at one ($\widehat{\pi}_2 = 0.14$). The CMP and ZICMP models estimates correctly neither the rate λ nor the dispersion ν. The ZkICMP model captures the inflation at zero($\widehat{\pi}_1 = 0.49$), at $k = 1$ ($\widehat{\pi}_2 = 0.20$) and underdispersion ($\widehat{\nu} = 2.46$). The estimated values are close to the true values. However, the estimate of the rate parameter ($\widehat{\lambda} = 17.56$) is not close to the true value of the parameter ($\lambda = 8$) and it has a high standard error. The ZkICMP model has AIC 1430.32 which is slightly less than the AIC values 1437.60 and 1437.65, respectively, of ZkINB and ZkIP models. Pairwise comparisons using the likelihood ratio tests for the simulated data reveal (1) CMP model fits better than the Poisson, (2) ZICMP fits better than ZIP, and (3) ZkICMP fits better than ZICMP model. Thus using LRT we conclude the CMP models fit significantly better than their Poisson counterparts, and among the CMP models it is the doubly inflated model, ZkICMP that beats the other models as expected. The results from AIC criterion concur with the LRT.

We also notice the results of ZkINB and ZkIP models are comparable. In the ZkINB model the estimates, standard errors, log likelihood and AIC are similar to the ZkIP model. Both the ZkIP and ZkINB models capture the inflations at zero and k, but they fail to capture the underlying underdispersion in the data. We obtain

similar results when we increase the sample size to 1000 and 2000. Notably, for large sample sizes not only the ZkICMP outperforms the other models, the parameter estimates get closer to the true values.

4.5.2 SIMULATION II

For the second simulated data we have used the parameter values $\lambda = 3$, $\nu = 1.5$, $\pi_1 = 0.4$ and $\pi_2 = 0.1$ for the ZkICMP distribution. The simulated data is underdispersed since $\nu > 1$ and inflated at zero ($\pi_1 = 0.4$) and at $k = 2$ ($\pi_2 = 0.1$). Unlike the previous example, we have simulated data for small sample sizes. But we did not see inflated frequencies until the sample size increased to about 200, and thus we have done our simulations for sample sizes $n = 200$, 500, 1000 and 2000. The results of our analysis are summarized in Table 14.

The estimate of λ is far from the true value for all sample sizes and for all models other than the ZkICMP. For the ZkICMP, the estimate of λ get closer to the true value as the sample size increases. The two models ZkIP and ZkINB capture both the inflations at zero and at 2 even for a sample of size 200. The estimates of π_1 and π_2 for these models are close to the true values. However, ZkINB model fails to capture the dispersion, for all sample sizes the estimate $\widehat{r}$ is approximately 0. The CMP model gives incorrect estimates for both the rate and dispersion parameters λ and ν for all sample sizes. The estimates do not seem to get closer to the true values even if the sample size increases. The zero inflated extension of the CMP model, ZICMP is able to capture the inflation at zero even for a small sample size, $\widehat{\pi}_1 = 0.3924$ is very close to the true value 0.4. However the estimates of $\widehat{\lambda} = 7.6011$ and $\widehat{\nu} = 2.4206$ are far from the true values of 3 and 1.5 respectively. The estimates of all the parameters for the ZkICMP model are close to their true values even for a small size. The AIC value is also the least for the ZkICMP model for any sample size.

The results from likelihood ratio test for pairwise comparisons between the models are given in Tables 16 and 17. These tables show results for all sample sizes, we can see ZkICMP is better than ZICMP, which is better than CMP. Similarly, ZkIP is better than ZIP, which is better than Poisson for all sample sizes. The goodness of fit results are displayed in Table 15. The Poisson model fits the worst for all sample

sizes. We observe the ZkIP and ZkICMP models fit equally well for $n = 200$ and 500. But as we increase the sample size the inflated frequencies increase leading to the increment in the error in the frequencies for the ZkIP model. This happens because the ZkIP model does account for the inflation at zero and $k = 2$ correctly but fails to capture the underdispersion which is captured by the ZkICMP model.

In summary the Poisson model is obviously the wrong model as it is unable to capture the inflations and underdispersion in the data. The ZIP model fails to account for the inflation at count 2 but not the inflation at 0. In most cases ZkIP does capture the inflation at zero and k but the rate parameter λ is estimated incorrectly. The ZkINB fails to capture the underdispersion in the data. In some cases, the CMP model does estimate accurately the rate and dispersion parameters. But the AIC of CMP model is higher than that of ZICMP. The ZICMP model allows the flexibility of capturing inflation at zero along with underdispersion. Finally we can easily conclude that the ZkICMP model outperform the other competing models. It not only captures the inflations at zero and k but also the underdispersion in the simulated data sets. The AIC turns out to be the be minimum for the ZkICMP model.

4.6 EXAMPLES

In this section we illustrate the application of the zero and k inflated Conway-Maxwell-Poisson (ZkICMP) model to analyze two real life data. The first data is an example where zero and one are inflated, and the second example has inflated frequencies for zero and count $k = 5$. Both the data were obtained from the National Health and Nutrition Examination Survey (NHANES). The NHANES has been collecting data related to the health and nutrition of children and adults in USA since early 1960s. And since 1999, it has been collecting annually, demographic, dietary information, and laboratory data of the sampled subjects.

4.6.1 DRUGS DATA

In this example the response variable is the number of joints/pipes of a drug smoked by the adults without a prescription from a doctor. Subjects aged between

18 and 59, were asked two questions in a survey. The first question was 'Have you ever used marijuana or hashish?'. If the answer was negative, the value of response variable is taken as zero. If it was positive, a follow up question 'How many joints/pipes did you smoke in a day?' was asked and the response was recorded. There are four covariates included in the survey. They are BMI, age, gender and family income as measured by the ratio to poverty level. More than 4000 people were surveyed but complete data was available for only 2481 subjects. The mean and variance of the count response were 0.70 and 1.20 respectively. The percentage of people who never smoked was 64.05%, and among the adults who smoked, 15.12% did so on an average of one joint, and 20.83% smoked more than a joint per day. Clearly, counts zero and one have high frequencies. Therefore this data is a perfect candidate for the models that we have been studying in this dissertation.

We fit the count models ZkICMP, ZICMP, CMP, ZkINB, ZkIP, ZIP and Poisson for this drug usage data. The parameter estimates and standard errors of ZkICMP, ZICMP models were computed in R using the non-linear optimization methods mentioned in Section 4.3.1. While, the results of CMP and Poisson models were obtained using SAS software count regression (Countreg) and Generalized linear model (GENMOD) procedures respectively. The results of ZkINB, ZkIP and ZIP models were obtained from finite mixture model (FMM) procedure in SAS. The maximized log-likelihood function and AIC values to compare the models were obtained as described in Section 4.4.3. The parameter estimates and standard errors for the models are displayed in Table 18.

The variance of the negative binomial (NB) distribution, after proper reparametrization, can be written as $\lambda + r\lambda^2$. If $r = 0$ then the NB has equidispersion similar to Poisson distribution. From Table 18, we can see that the results are similar for the three models ZkINB, ZkIP and ZIP because $\widehat{r} = 0$, and also because $\widehat{\delta} = -0.8456$ which indicates the estimate of π_2 is small. The AIC values are also similar for the three ZkINB, ZkIP and ZIP models. Here, it is irrelevant to perform an LRT to compare ZkIP to ZIP model as the value of test statistic will be zero. Theoretically, ZkINB is not nested within ZkIP or ZIP model so we do not perform a test to compare ZkINB to ZkIP or ZIP models. Thus ZIP is preferable among these three models since it has the least number of parameters.

For testing $H_0 : \nu = 1$, the LRT test in (43) from Section 4.4.3 gives $-2\log\Lambda =$

21.66, which has a p-value that is less than 0.0001. Thus we reject $H_0 : \nu = 1$ and conclude that ZkICMP model is significantly better than ZkIP. Similarly, LRT test also shows that ZICMP fits better than ZIP. The countreg procedure in SAS gives $\widehat{\nu} = 0.0019$ for the CMP model, but it gave a singular Hessian matrix and we could not get the standard error. The estimate of the dispersion parameter of the ZICMP model is $\widehat{\nu} = 1.50$ while, in ZkICMP model it is $\widehat{\nu} = 3.84$ and both values are statistically significant. This indicates existence of significant underdispersion in both the models.

We also performed LRT test as described in Section 4.4.3 to test for the inflation probabilities. The LRT test statistic is $-2\log\Lambda = 180.34$ with a $p-$value< 0.0001. It implies that the ZICMP model is significantly better than the CMP model. Also, the inflation at zero is significant. The LRT also showed π_2 is significant, thus ZkICMP is better than ZICMP $(-2\log\Lambda = 10.82, p - \text{value} = 0.0005)$. It is not possible to use LRT to compare ZkINB and ZkIP model as they are not nested. Similarly, we cannot compare ZkINB and ZIP model using the LRT criterion. While, here the ZkIP reduces to ZIP so the test of $\pi_2 = 0$ is not required. Comparing ZIP to Poisson we get, $-2\log\Lambda = 563.42, p - \text{value} < 0.0001$. Thus ZIP model fits better than the Poisson model.

Table 18, also includes the AIC values of the models. According to the AIC criterion, the CMP models perform better than their Poisson counterparts. Further, we see the inflated models perform better than the standard models, both CMP and Poisson. The smallest AIC values are for ZkICMP, ZICMP, CMP and ZIP models. Using the LRT and AIC criterions, the ZIP performs better than the Poisson, ZkIP and ZkINB models. Also, we see that the ZkICMP and ZICMP models are better than CMP model.

To select the best fit model we further analyze the residuals and compare the fitted and observed frequencies. The observed and expected frequencies for the comparable models are shown in the Table 19. The ZIP model does not give a good fit to the data. The predicted frequencies of zero and one in the ZICMP model are close to the observed values but ZkICMP outperforms it by predicting closer values for all the count values. Figure 8 shows that the closest frequencies are predicted by ZkICMP model.

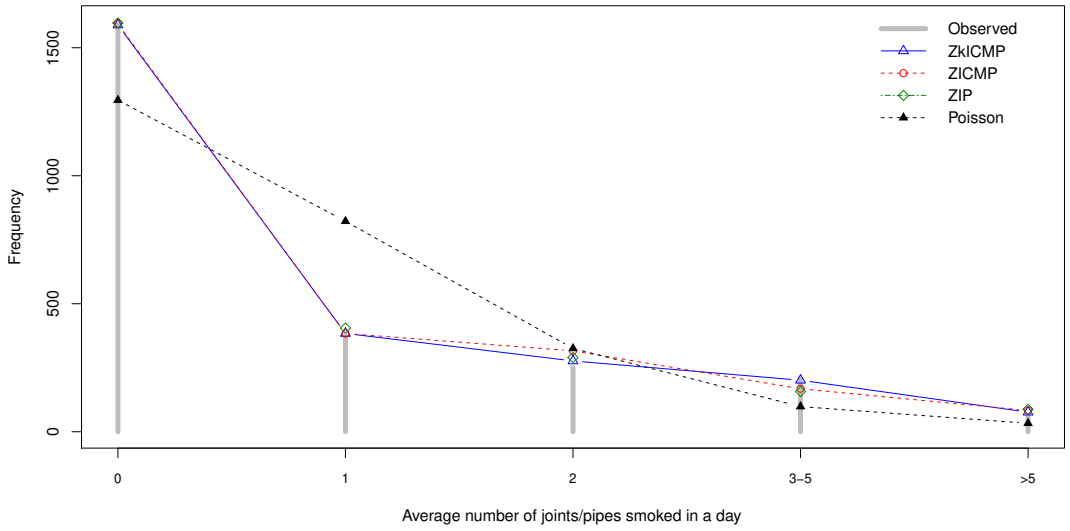
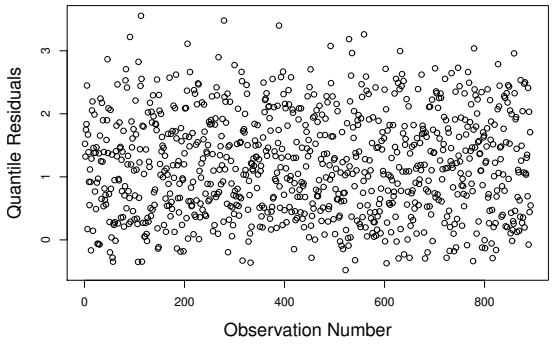


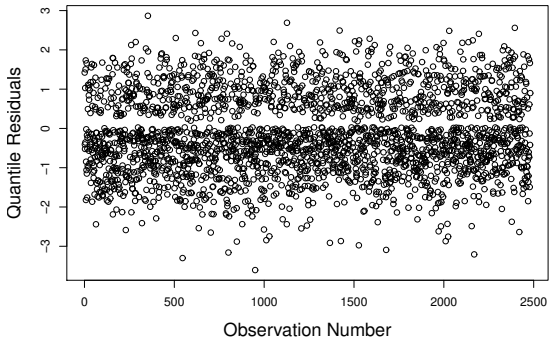
Figure 8. Observed and Expected Frequencies for drugs data.

The residual analysis is performed for the comparable models, ZIP, ZICMP, ZkICMP and Poisson. Figure 9 shows that the residuals of Poisson model have a pattern. They are concentrated in between zero and three. The QQ plot in Figure 10 clearly shows deviation from a straight line. Hence, Poisson does not give a good fit to the data. The residual plots of ZIP model are not concentrated around zero. This is further supported by the QQ plot. Thus, the residuals of ZIP are not from standard normal. The residual plot of ZICMP is nearly random with some deviation in the QQ plot. Here, the best model is ZkICMP model as the residual plot looks completely random and the QQ plot has most of the quantiles agreeing with the standard normal quantiles apart from some deviation which might be due to few outlier observations.

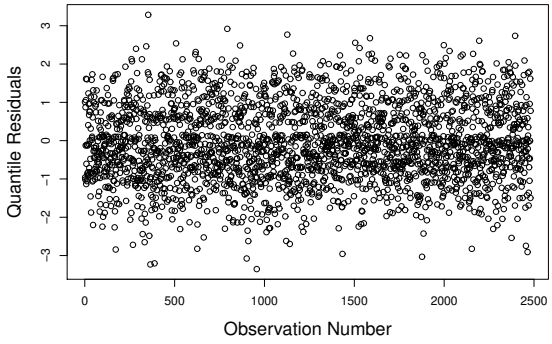
In summary, for this data the empirical mean (0.70) is slightly lower than variance (1.20). But there is an underlying underdispersion in the data which is overtaken by the excess number of zeros. There is also a significant peak at $k = 1$. The standard Poisson and negative binomial have the ability to capture only over or equidispersion. They lack the ability to account for the underdispersion. The ZkICMP model



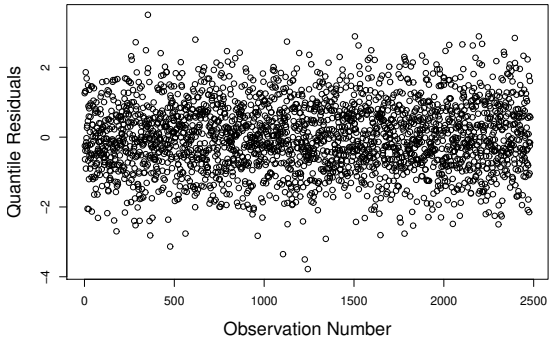
(a) Residual plot of Poisson



(b) Residual plot of ZIP

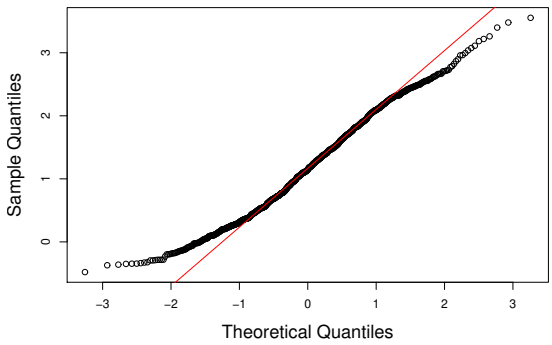


(c) Residual plot of ZICMP

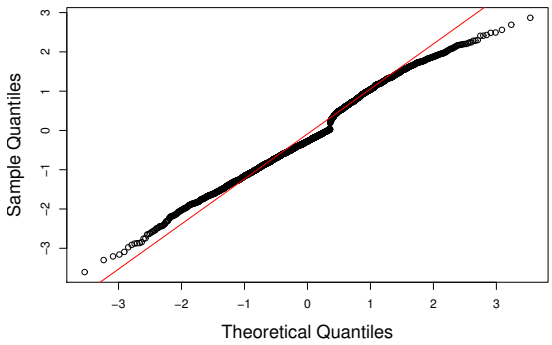


(d) Residual plot of ZkICMP

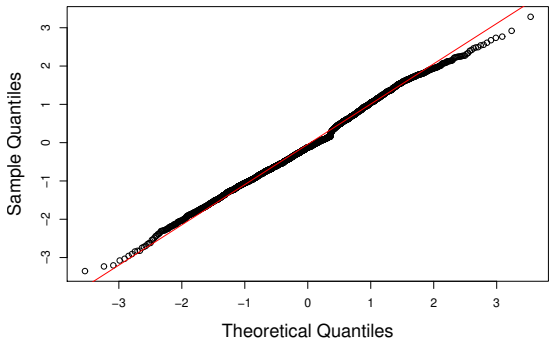
Figure 9. Randomized quantile residual plots for drugs data.



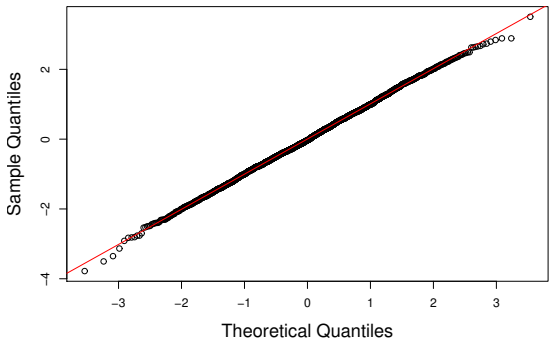
(a) QQ plot of Poisson



(b) QQ plot of ZIP



(c) QQ plot of ZICMP



(d) QQ plot of ZkICMP

Figure 10. QQ plots for drugs data.

successfully captures not only the inflated frequencies at zero and k but also the underdispersion in the data. It turns out to be the best model for the observed data.

4.6.2 EXERCISE DATA

The data for this example were also obtained from NHANES. In this example the response variable is number of times a subject did vigorous or moderate activities in a week. The variable is constructed using the following four questions on the questionnaire survey: (1) Have you done vigorous activity during the week? (2) Have you done moderate activity during the week? (3) How many days did you do vigorous activity in the week? and (4) How many days did you do moderate activity during the week? The response variable is taken as zero if the answer to the first two questions is negative. Otherwise the response variable is taken as the sum of the values obtained for questions three and four.

There were several relevant covariates that we could take into consideration. However, the data were incomplete for several covariates. We have selected the variables where complete data was available and they are age, body mass index (BMI), body weight, ratio of family income to poverty (ratio), gender, average systolic blood pressure (BP) and average diastolic blood pressure (BP). The variables age, ratio and gender were in the demographic file of the survey. While, BMI and weight were saved as 'BMXWT' and 'BMXBMI' respectively in the body measure file. The average BP data was obtained by averaging the four readings of the examined subjects. The readings were in the blood pressure file under the examination data section.

The respondents of both genders included in the study were between 12 to 80 years old. The total number of subjects included in the data were 6122. In the data 62.15% never did any activity, while 7.87% subjects did either vigorous or moderate or both of the exercises five times a week. Thus zero and count 5 occur with high frequencies in the data. The range of the responses varied between zero and 13. The sample mean and variance were 2.03 and 10.18 respectively. Clearly, the data is overdispersed. Also, the observed frequencies of the counts zero and five are more than that expected under a Poisson regression model.

We checked the correlations between the covariates and found a significant high (0.9) correlation between BMI and weight, and made a decision to drop weight from

further analysis. As in the previous example, we fit various models to the data. The results are given in Table 20. The table shows the covariates age and average systolic BP are insignificant for ZIP, ZkIP, ZkINB, ZICMP and ZkICMP models. We refit all the models removing these covariates and including only BMI, ratio, gender and average diastolic BP. The parameter estimates and standard errors along with the loglikelihood and AIC of the models are given in Table 21.

We observe, the estimate $\widehat{\nu}$ for the CMP model is close to zero and SAS software failed to give a standard error. For the ZICMP and ZkICMP models, the estimate of ν is less than 1 indicating the presence of overdispersion in the data. The estimate $\widehat{r} = 0.2108$ for the ZkINB model is significant which also supports the existence of overdispersion in the data. For all the models the covariates exhibit similar relation with the number of times a subject did activity/activities in a week. The covariates BMI, gender (male) and average diastolic blood pressure have positive relation with the response variable whereas the variable ratio of family income to poverty has a negative relation.

Table 21 also shows the AIC values of CMP models are smaller than their Poisson analogs for both the inflated (single and double) and non-inflated cases. The LRT tests also show the CMP models are significantly different than their corresponding Poisson models. Thus the CMP models are better than their Poisson counterparts as they have the ability to capture underlying overdispersion in the data. From Table 21, we can see the AIC value of the inflated models is less than that of the standard CMP and classical Poisson model. A comparison between observed and expected frequencies is given in Table 22. The expected frequencies from the ZkICMP model have ABE of 614.79 and a chi-square value of 434.72, and these values are the smallest among the competing models. Thus the ZkICMP model fits the data best, which can also be seen from the graph in Figure 11.

For a postmortem analysis we plotted the randomized residuals as described in Sellers and Raim (2016). If the model fits the data correctly the plots in Figure 12 should exhibit a random behavior. We observe, the residuals for the ZkIP model do not appear completely random. However, the residuals plots of other models, ZkINB, ZICMP and ZkICMP appear to be random. We also notice that most of the residuals lie between -3 and 3. Figure 13 shows the QQ plots for the ZkIP, ZkINB, ZICMP and ZkICMP models. The sample quartiles of the ZkIP model do not agree with its

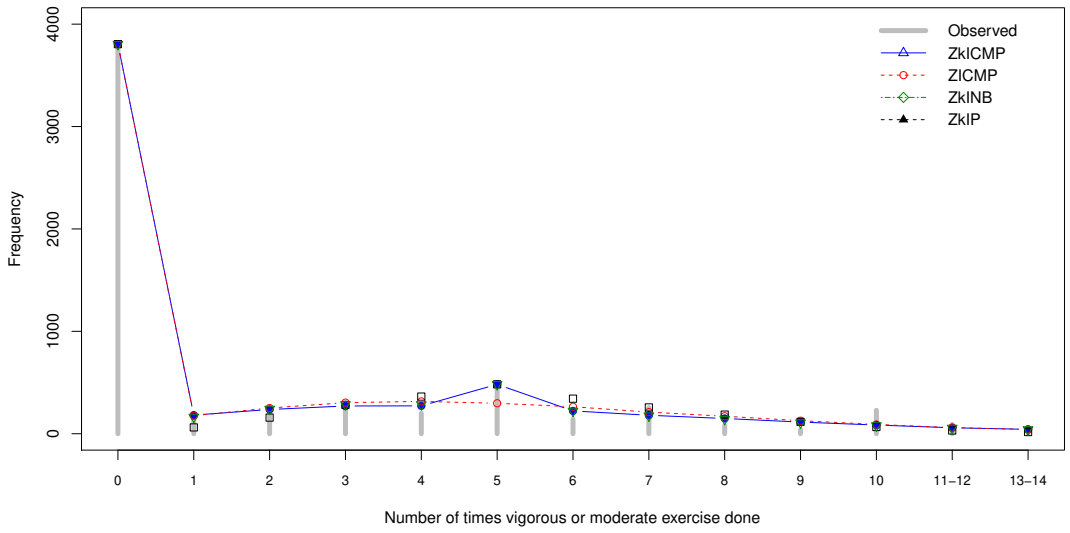


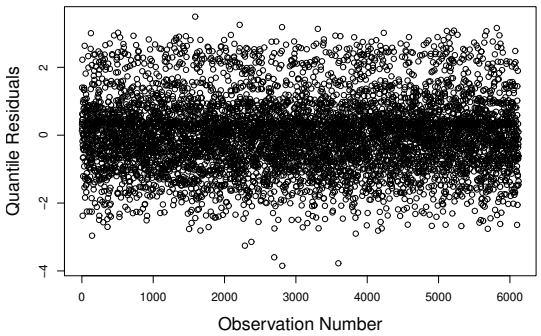
Figure 11. Observed and Expected Frequencies for exercise data.

theoretical quartiles. The plot of ZICMP model does not provide good comparison of the quartiles in the lower and upper tail. The plots of ZkINB and ZkICMP models are mostly comparable except from some difference in the tails. The QQ plot of ZkICMP gives a better fit apart from a small deviation from the straight line.

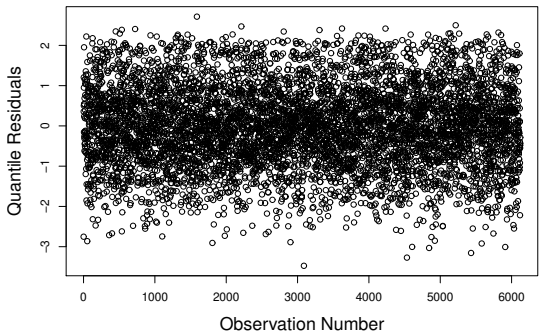
We conclude for this data, the models capturing overdispersion and double inflation would be the most appropriate. Using the AIC criterion, ZkINB model performs better than ZICMP. But it is the ZkICMP model that captures not only the peaks at zero and 5 but also the dispersion in the data. The ZkICMP provides the best fit to the observed data here.

4.7 SUMMARY

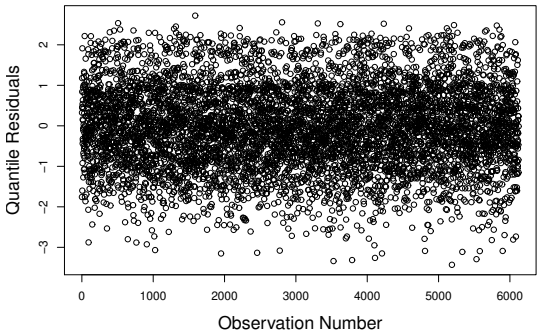
In summary, in this chapter we introduce a new regression model to capture inflation at zero and a count value k and model under (over) dispersion in the data. We refer to it as zero and k inflated Conway-Maxwell-Poisson (ZkICMP) model. The model is an extension of ZICMP model. The special cases of ZkICMP model are ZkIP,



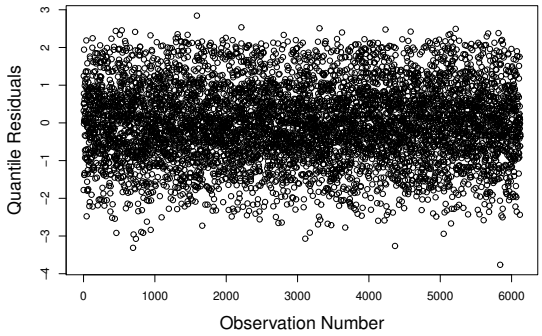
(a) Residual plot of ZkIP



(b) Residual plot of ZkINB

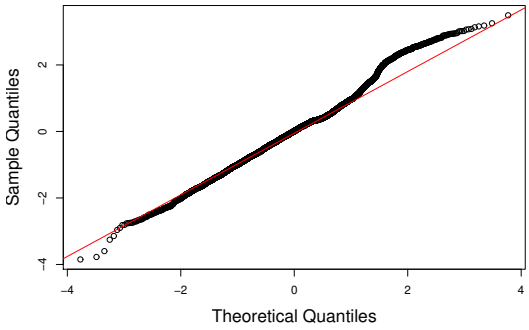


(c) Residual plot of ZICMP

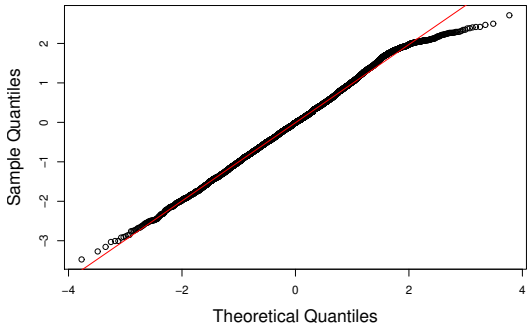


(d) Residual plot of ZkICMP

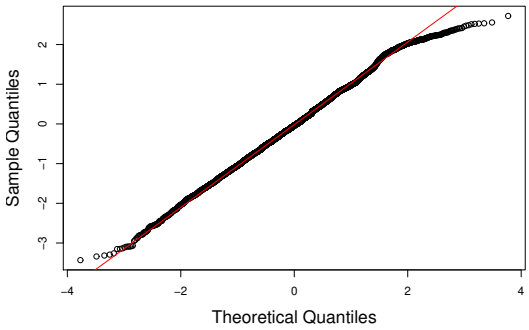
Figure 12. Randomized quantile residual plots for exercise data.



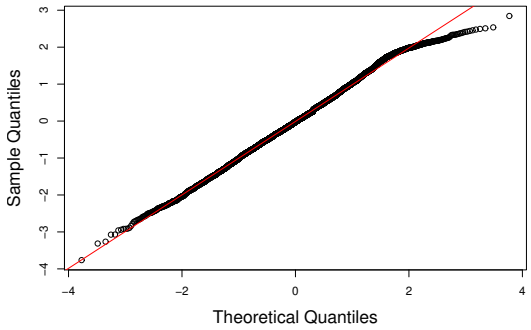
(a) QQ plot of ZkIP



(b) QQ plot of ZkINB



(c) QQ plot of ZICMP



(d) QQ plot of ZkICMP

Figure 13. QQ plots for exercise data.

ZkIG. The advantage of ZkICMP model over ZkINB model is that ZkINB can only capture overdispersion but ZkICMP provides the flexibility to capture underdispersion as well. Note, both models have the same number of parameters, but ZkICMP provides an extra feature without any extra cost. The ZkICMP model is easy to construct and understand. In this chapter we assumed only the rate parameter λ depends on the covariates. Possible extensions of this model are straightforward. We could for example let $\log(\pi_1/\pi_3) = \mathbf{x_i^T}\,\boldsymbol{\gamma}$ and $\log(\pi_2/\pi_3) = \mathbf{x_i^T}\,\boldsymbol{\delta}$ and study the dependence of the inflations on the covariates. Few more extensions of ZkICMP model similar to Shmueli et al. (2005) and Sellers and Raim (2016) could also be pursued.

Table 12. Estimates and standard errors for simulated data I

n	Parameter	ZkICMP	ZICMP	CMP	ZkINB	ZkIP	ZIP	Poisson
2000	$\widehat{\lambda}$	9.2069	0.8597	0.5302	2.0903	2.0903	1.6099	0.9960
		(3.5742)	(0.1001)	(0.0193)	(0.0829)	(0.0829)	(0.0467)	(0.0223)
	$\widehat{\nu}$	2.0851	0.4412	0.0699	–	–	–	–
		(0.2874)	(0.0976)	(0.0385)				
	$\widehat{\pi}_1$	0.5001	0.2469	–	0.4540	0.4540	0.3813	–
		(0.0187)	(0.0408)		(0.0144)	(0.0144)	(0.0157)	
	$\widehat{\pi}_2$	0.1948	–	–	0.1333	0.1333	–	–
		(0.0198)			(0.0082)	(0.0083)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-2726.83	-2754.49	-2765.30	-2735.25	-2735.27	-2768.28	-2957.60
	AIC	5461.67	5514.98	5534.61	5476.50	5476.54	5540.55	5917.19
1000	$\widehat{\lambda}$	7.9076	0.8371	0.5665	1.8714	1.8715	1.4685	0.9730
		(4.7587)	(0.1469)	(0.0303)	(0.1143)	(0.1143)	(0.0629)	(0.0312)
	$\widehat{\nu}$	2.0774	0.4825	0.1682	–	–	–	–
		(0.4549)	(0.1527)	(0.0609)				
	$\widehat{\pi}_1$	0.4826	0.1992	–	0.4205	0.4205	0.3374	–
		(0.0318)	(0.0662)		(0.0239)	(0.0239)	(0.0240)	
	$\widehat{\pi}_2$	0.1992	–	–	0.1280	0.1280	–	–
		(0.0350)			(0.0144)	(0.0144)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-1350.23	-1360.17	-1363.16	-1353.50	-1353.52	-1365.12	-1430.21
	AIC	2708.46	2726.34	2730.32	2713.0	2713.04	2734.23	2862.42
500	$\widehat{\lambda}$	17.5561	1.0181	0.5532	2.3219	2.3220	1.8048	1.1060
		(12.7659)	(0.2217)	(0.0387)	(0.1621)	(0.1621)	(0.0956)	(0.0470)
	$\widehat{\nu}$	2.4564	0.5076	0.0556	–	–	–	–
		(0.5315)	(0.1773)	(0.0691)				
	$\widehat{\pi}_1$	0.4864	0.2868	–	0.4470	0.4470	0.3872	–
		(0.0294)	(0.0639)		(0.0257)	(0.0257)	(0.0291)	
	$\widehat{\pi}_2$	0.2021	–	–	0.1347	0.1347	–	–
		(0.0292)			(0.0150)	(0.0150)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-711.16	-723.06	-728.21	-715.80	-715.82	-726.37	-787.73
	AIC	1430.32	1452.12	1460.42	1437.60	1437.65	1456.74	1577.46
200	$\widehat{\lambda}$	5.4587	1.1827	0.5247	1.8256	1.8255	1.5936	0.9600
		(6.9954)	(0.4907)	(0.0614)	(0.2475)	(0.2476)	(0.1493)	(0.0693)
	$\widehat{\nu}$	1.8191	0.7289	0.0819	–	–	–	–
		(0.9656)	(0.3632)	(0.1274)				
	$\widehat{\pi}_1$	0.5056	0.3467	–	0.4416	0.4416	0.3976	–
		(0.0861)	(0.1000)		(0.0608)	(0.0608)	(0.0499)	
	$\widehat{\pi}_2$	0.1365	–	–	0.0719	0.0719	–	–
		(0.0870)			(0.0306)	(0.0306)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-267.97	-268.93	-271.41	-268.40	-268.38	-269.19	-288.85
	AIC	543.95	543.87	546.83	542.80	542.76	542.39	579.71

NOTE: Standard errors are given in parenthesis. Here $k = 1$ and the generating values of the parameters are $\lambda = 8, \nu = 2, \pi_1 = 0.5, \pi_2 = 0.2$.

Table 13. Frequency comparisons for simulated data I

n	Count	Observed	ZkICMP	ZICMP	ZkIP	Poisson
2000	0	1010	1010.00	1010.00	1010.00	738.71
	1	480	480.00	443.77	213.33	735.75
	2	198	196.10	281.00	222.97	366.40
	3	177	182.70	148.79	155.36	121.65
	4	98	93.43	69.40	81.19	30.29
	5	31	30.00	29.33	33.94	6.03
	6	4	6.59	11.44	11.82	1.00
	7	2	1.05	4.17	3.53	0.14
	ABE		16.71	187.32	342.39	848.34
	χ^2		2.33	50.67	348.73	578.97
1000	0	490	490.00	490.00	490.00	377.95
	1	258	258.00	243.44	258.00	367.74
	2	112	110.17	145.86	121.69	178.91
	3	83	88.91	71.87	75.91	58.03
	4	46	39.47	30.82	35.52	14.11
	5	8	11.02	11.87	13.29	2.75
	6	3	2.11	4.18	4.15	0.45
	ABE		18.18	79.79	33.70	353.37
	χ^2		2.71	19.53	6.95	198.47
500	0	244	244.00	244.00	244.00	165.44
	1	115	115.00	102.41	115.00	182.98
	2	47	44.62	73.34	55.30	101.19
	3	46	52.71	42.75	42.80	37.30
	4	36	30.72	21.54	24.85	10.31
	5	10	10.35	9.69	11.54	2.28
	6	2	2.23	3.97	4.47	0.42
	ABE		14.95	58.93	26.65	244.40
	χ^2		1.92	21.96	8.06	189.61
200	0	104	104.00	104.00	104.00	76.58
	1	23	24.29	29.25	26.12	35.29
	2	22	17.97	15.53	15.90	11.29
	3	4	7.88	6.69	7.25	2.71
	4	3	2.30	2.45	2.65	0.52
	5	1	0.48	0.78	0.81	0.08
	ABE		10.41	18.18	13.03	85.62
	χ^2		3.65	5.39	4.27	59.44

NOTE: ABE is absolute error. Here $k = 1$.

Table 14. Estimates and standard errors for simulated data II

n	Parameter	ZkICMP	ZICMP	CMP	ZkINB	ZkIP	ZIP	Poisson
2000	$\widehat{\lambda}$	3.3416	7.6842	0.7449	1.8049	1.8049	1.7538	1.1750
		(0.6286)	(1.2603)	(0.0285)	(0.0531)	(0.0531)	(0.0455)	(0.0242)
	$\widehat{\nu}$	1.5633	2.3812	0.3642	–	–	–	–
		(0.1734)	(0.1559)	(0.0429)				
	$\widehat{\pi}_1$	0.4040	0.4259	–	0.3637	0.3637	0.3300	–
		(0.0157)	(0.0122)		(0.0130)	(0.0130)	(0.0149)	
	$\widehat{\pi}_2$	0.1114	–	–	0.1359	0.1359	–	–
		(0.0161)			(0.0082)	(0.0082)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-2788.86	-2811.44	-2957.01	-2794.85	-2794.87	-2860.37	-3044.38
	AIC	5585.72	5628.87	5918.01	5595.70	5595.73	5724.73	6090.77
1000	$\widehat{\lambda}$	2.8471	7.4459	0.7781	1.5904	1.5905	1.5912	1.1230
		(0.8582)	(1.8122)	(0.0434)	(0.0706)	(0.0706)	(0.0617)	(0.0335)
	$\widehat{\nu}$	1.5488	2.4968	0.4616	–	–	–	–
		(0.2862)	(0.2407)	(0.0667)				
	$\widehat{\pi}_1$	0.3788	0.4138	–	0.3282	0.3282	0.2943	–
		(0.0255)	(0.0178)		(0.0197)	(0.0197)	(0.0226)	
	$\widehat{\pi}_2$	0.1080	–	–	0.1331	0.1331	–	–
		(0.0234)			(0.0124)	(0.0124)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-1369.53	-1378.40	-1437.91	-1371.60	-1371.62	-1402.54	-1465.19
	AIC	2747.06	2762.81	2879.82	2749.20	2749.23	2809.07	2932.38
500	$\widehat{\lambda}$	3.7276	7.0113	0.7215	1.8874	1.8873	1.8242	1.1920
		(1.3399)	(2.2265)	(0.0542)	(0.1071)	(0.1071)	(0.0927)	(0.0488)
	$\widehat{\nu}$	1.6161	2.2362	0.3133	–	–	–	–
		(0.3284)	(0.2956)	(0.0818)				
	$\widehat{\pi}_1$	0.4160	0.4310	–	0.3756	0.3756	0.3466	–
		(0.0300)	(0.0246)		(0.0261)	(0.0261)	(0.0290)	
	$\widehat{\pi}_2$	0.0929	–	–	0.1199	0.1199	–	–
		(0.0316)			(0.0159)	(0.0159)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-705.10	-709.33	-746.68	-707.10	-707.12	-720.10	-773.82
	AIC	1418.19	1424.67	1497.36	1420.20	1420.24	1444.20	1549.64
200	$\widehat{\lambda}$	2.8648	7.6011	0.8323	1.7307	1.7307	1.6960	1.2150
		(1.7051)	(3.9083)	(0.1029)	(0.1618)	(0.1618)	(0.1383)	(0.0779)
	$\widehat{\nu}$	1.4642	2.4206	0.4751	–	–	–	–
		(0.5506)	(0.4956)	(0.1419)				
	$\widehat{\pi}_1$	0.3607	0.3924	–	0.3213	0.3213	0.2836	–
		(0.0538)	(0.0388)		(0.0409)	(0.0409)	(0.0481)	
	$\widehat{\pi}_2$	0.1284	–	–	0.1496	0.1496	–	–
		(0.0514)			(0.0287)	(0.0287)		
	$\widehat{r}$	–	–	–	< 0.0001	–	–	–
					–			
	$\log L_{obs}$	-282.10	-284.86	-297.85	-282.50	-282.49	-289.98	-303.58
	AIC	572.19	575.71	599.71	571.00	570.99	583.96	609.15

NOTE: Standard errors are given in parenthesis. Here $k = 2$ and true values of the parameters are $\lambda = 3, \nu = 1.5, \pi_1 = 0.4, \pi_2 = 0.1$.

Table 15. Frequency comparisons for simulated data II

n	Count	Observed	ZkICMP	ZICMP	ZkIP	Poisson
2000	0	892	892.00	892.00	892.00	617.64
	1	281	280.54	309.37	297.10	725.72
	2	540	540.00	456.30	540.00	426.36
	3	188	190.29	256.29	161.32	166.99
	4	75	72.80	72.56	72.79	49.05
	5	19	19.65	12.08	26.28	11.53
	6	5	3.99	1.30	7.90	2.26
	ABE		6.60	193.41	55.18	889.89
	χ^2		0.37	50.71	8.44	449.23
1000	0	438	438.00	438.00	438.00	325.30
	1	169	168.49	180.08	174.62	365.31
	2	272	272.00	237.56	272.00	205.12
	3	82	85.15	113.87	73.62	76.78
	4	32	28.32	26.61	29.27	21.56
	5	6	6.67	3.56	9.31	4.84
	6	1	1.18	0.30	2.47	0.91
	ABE		8.19	85.91	21.51	392.80
	χ^2		0.69	18.96	3.44	172.04
500	0	226	226.00	226.00	226.00	151.81
	1	68	67.05	73.52	72.12	180.95
	2	128	128.00	109.41	128.00	107.85
	3	46	51.49	65.76	42.82	42.85
	4	26	20.43	20.77	20.20	12.77
	5	6	5.65	3.98	7.63	3.04
	ABE		12.36	51.12	14.73	226.63
	χ^2		2.14	11.85	2.48	127.34
200	0	83	83.00	83.00	83.00	59.34
	1	31	31.13	34.31	32.45	72.10
	2	58	58.00	48.72	58.00	43.80
	3	20	18.53	25.92	16.20	17.74
	4	4	6.98	6.87	7.01	5.39
	5	4	1.89	1.06	2.43	1.31
	ABE		6.67	24.33	9.83	85.30
	χ^2		3.73	12.77	3.27	43.64

NOTE: ABE is absolute error. Here $k = 2$.

Table 16. Testing zero inflation for simulated data II

n		ZICMP *v/s CMP*	ZIP *v/s Poisson*
2000	Likelihood Ratio	291.14	368.02
	asy. dist.	$0.5\chi^2_0 + 0.5\chi^2_1$	
	p-value	< 0.0001	< 0.0001
1000	Likelihood Ratio	119.02	125.30
	asy. dist.	$0.5\chi^2_0 + 0.5\chi^2_1$	
	p-value	< 0.0001	< 0.0001
500	Likelihood Ratio	74.70	107.44
	asy. dist.	$0.5\chi^2_0 + 0.5\chi^2_1$	
	p-value	< 0.0001	< 0.0001
200	Likelihood Ratio	25.98	27.20
	asy. dist.	$0.5\chi^2_0 + 0.5\chi^2_1$	
	p-value	< 0.0001	< 0.0001

NOTE: The test is $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 > 0$ at $\alpha = 0.05$.

Table 17. Testing k inflation for simulated data II

n		ZkICMP *v/s ZICMP*	ZkIP *v/s ZIP*
2000	Likelihood Ratio	45.16	851.44
	asy. dist.	$0.5\chi_0^2 + 0.5\chi_1^2$	
	p-value	< 0.0001	< 0.0001
1000	Likelihood Ratio	17.74	392.36
	asy. dist.	$0.5\chi_0^2 + 0.5\chi_1^2$	
	p-value	< 0.0001	< 0.0001
500	Likelihood Ratio	8.46	192.46
	asy. dist.	$0.5\chi_0^2 + 0.5\chi_1^2$	
	p-value	0.0018	< 0.0001
200	Likelihood Ratio	5.52	27.20
	asy. dist.	$0.5\chi_0^2 + 0.5\chi_1^2$	
	p-value	0.0094	< 0.0001

NOTE: The test is $H_0 : \pi_2 = 0$ against $H_1 : \pi_2 > 0$ at $\alpha = 0.05$. Here $k = 2$.

Table 18. Estimates and standard errors for drugs data

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZkIP	ZIP	Poisson
Intercept	4.5274*	1.1522*	-0.7962*	0.5678*	0.5678*	0.5678*	-0.1647*
	(0.7977)	(0.2196)	(0.0635)	(0.0984)	(0.0984)	(0.0984)	(0.0834)
Age	-0.0163*	-0.0093*	-0.0059*	-0.0081*	-0.0081*	-0.0081*	-0.0103*
	(0.0050)	(0.0028)	(0.0015)	(0.0024)	(0.0024)	(0.0024)	(0.0020)
Income	-0.1854*	-0.1038*	-0.0368*	-0.0787*	-0.0787*	-0.0787*	-0.0637*
	(0.0390)	(0.0224)	(0.0120)	(0.0181)	(0.0181)	(0.0181)	(0.0157)
Gender	0.4945*	0.4881*	0.3357*	0.4444*	0.4444*	0.4444*	0.5675*
	(0.1223)	(0.0719)	(0.0391)	(0.0609)	(0.0609)	(0.0609)	(0.0501)
$\widehat{\gamma}$	0.9279	0.2562	–	0.0447	0.0447	0.0447	–
	(0.0745)	(0.0738)		(0.0609)	(0.0609)	(0.0609)	
$\widehat{\delta}$	-0.8456	–	–	-16.1765	-16.5838	–	–
	(0.1867)			(377.00)	(462.14)		
$\widehat{\nu}$	3.8416*	1.4968*	0.0019	–	–	–	–
	(0.5674)	(0.1608)	–				
$\widehat{r}$	–	–	–	0.0000	–	–	–
				–			
$\widehat{\pi}_1$	0.6389	0.5637	–	0.5112	0.5112	0.5112	–
	(0.0172)	(0.0182)		(0.0152)	(0.0152)	(0.0152)	
$\widehat{\pi}_2$	0.1084	–	–	0.0000	0.0000	–	–
	(0.0181)			(< 0.0001)	(< 0.0001)		
$\log L_{obs}$	-2709.42	-2714.83	-2805.00	-2720.25	-2720.25	-2720.25	-3001.96
AIC	5432.85	5441.66	5619.00	5452.50	5452.50	5450.50	6011.93

NOTE: Standard errors are given in parenthesis. Also, the significant regression and dispersion parameters at $\alpha = 0.05$ are marked with an asterisk.

Table 19. Frequency comparisons for drugs data

Count	Observed	ZkICMP	ZICMP	ZIP	Poisson
0	1589	1589.01	1593.65	1596.63	1295.34
1	375	383.99	383.89	405.12	822.19
2	250	276.54	316.05	290.04	325.35
$3-5$	206	201.14	167.99	157.15	98.24
>5	61	76.40	82.86	87.87	33.01
ABE		55.80	139.46	153.51	951.94
χ^2		5.98	28.39	31.21	469.17

Table 20. Estimates and standard errors for exercise data

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZkIP	ZIP	Poisson
Intercept	0.3419*	0.4814*	-0.6003*	1.2415*	1.3109*	1.3221*	0.0483
	(0.0668)	(0.0673)	(0.0399)	(0.1194)	(0.0772)	(0.0755)	(0.0726)
Age	-0.0006	-0.0006	-0.0026*	-0.0010	-0.0005	-0.0005	-0.0079*
	(0.0004)	(0.0004)	(0.0003)	(0.0009)	(0.0006)	(0.0006)	(0.0005)
BMI	0.0025*	0.0025*	0.0034*	0.0052*	0.0041*	0.0039*	0.0115*
	(0.0010)	(0.0010)	(0.0007)	(0.0022)	(0.0014)	(0.0013)	(0.0012)
Ratio	-0.0113*	-0.0112*	-0.0118*	-0.0229*	-0.0191*	-0.0183*	-0.0372*
	(0.0045)	(0.0045)	(0.0032)	(0.0092)	(0.0060)	(0.0059)	(0.0056)
Gender	0.0641*	0.0635*	0.1253*	0.1225*	0.0979*	0.0946*	0.3864*
	(0.0143)	(0.0144)	(0.0106)	(0.0296)	(0.0193)	(0.0188)	(0.0185)
Avg. Systolic	-0.0003	-0.0003	-0.0007*	-0.0006	-0.0003	-0.0003	-0.0016*
	(0.0005)	(0.0005)	(0.0004)	(0.0011)	(0.0007)	(0.0007)	(0.0007)
Avg. Diastolic	0.0027*	0.0027*	0.0036*	0.0052*	0.0044*	0.0042*	0.0108*
	(0.0006)	(0.0006)	(0.0004)	(0.0013)	(0.0008)	(0.0008)	(0.0008)
$\widehat{\gamma}$	0.5169	0.4301	–	0.5451	0.5321	0.4865	–
	(0.0298)	(0.0277)		(0.0291)	(0.0286)	(0.0265)	
$\widehat{\delta}$	-2.2577	–	–	-2.2708	-3.0736	–	–
	(0.1090)			(0.1129)	(0.2406)		
$\widehat{\nu}$	0.3799*	0.4588 *	0.0012	–	–	–	–
	(0.0220)	(0.0221)	–				
$\widehat{r}$	–	–	–	0.2114*	–	–	–
				(0.0163)			
$\widehat{\pi}_1$	0.6029	0.6059	–	0.6099	0.6194	0.6193	–
	(0.0071)	(0.0066)		(0.0063)	(0.0064)	(0.0062)	
$\widehat{\pi}_2$	0.0376	–	–	0.0365	0.0168	–	–
	(0.0039)			(0.0040)	(0.0015)		
$\log L_{obs}$	-9613.15	-9676.54	-11609.00	-9633.60	-9885.30	-9895.70	-17466.45
AIC	19246.30	19371.08	23234.00	19287.20	19788.60	19807.40	34946.90

NOTE: Standard errors are given in parenthesis. Also, the significant regression and dispersion parameters at $\alpha = 0.05$ are marked with an asterisk.

Table 21. Significant estimates and standard errors for exercise data

Parameters	ZkICMP	ZICMP	CMP	ZkINB	ZkIP	ZIP	Poisson
Intercept	0.3215*	0.4644*	-0.6683*	1.1956*	1.2856*	1.2993*	-0.1191*
	(0.0566)	(0.0567)	(0.0312)	(0.0915)	(0.0596)	(0.0582)	(0.0557)
BMI	0.0022*	0.0022*	0.0026*	0.0048*	0.0038*	0.0037*	0.0081*
	(0.0010)	(0.0010)	(0.0007)	(0.0021)	(0.0013)	(0.0013)	(0.0012)
Ratio	-0.0122*	-0.0121*	-0.0162*	-0.0243*	-0.0200*	-0.0191*	-0.0500*
	(0.0044)	(0.0045)	(0.0032)	(0.0091)	(0.0060)	(0.0058)	(0.0056)
Gender	0.0632*	0.0625*	0.1254*	0.1206*	0.0967*	0.0934*	0.3830*
	(0.0142)	(0.0142)	(0.0106)	(0.0291)	(0.0190)	(0.0186)	(0.0183)
Avg. Diastolic	0.0023*	0.0023*	0.0024*	0.0046*	0.0040*	0.0038*	0.0075*
	(0.0006)	(0.0006)	(0.0004)	(0.0011)	(0.0008)	(0.0007)	(0.0007)
$\widehat{\gamma}$	0.5191	0.4319	–	0.5460	0.5322	0.4866	–
	(0.0298)	(0.0276)		(0.0291)	(0.0285)	(0.0265)	
$\widehat{\delta}$	-2.2581	–	–	-2.2714	-3.0767	–	–
	(0.1092)			(0.1130)	(0.2413)		
$\widehat{\nu}$	0.3817*	0.4605*	0.0009	–	–	–	–
	(0.0220)	(0.0221)	–				
$\widehat{r}$	–	–	–	0.2108*	–	–	–
				(0.0162)			
$\widehat{\pi}_1$	0.6034	0.6063	–	0.6101	0.6194	0.6193	–
	(0.0071)	(0.0066)		(0.0063)	(0.0064)	(0.0062)	
$\widehat{\pi}_2$	0.0375	–	–	0.0365	0.0168	–	–
	(0.0039)			(0.0015)	(0.0015)		
$\log L_{obs}$	-9614.88	-9678.00	-11661.00	-9634.65	-9886.15	-9896.45	-17628.99
AIC	19245.75	19369.99	23335.00	19285.30	19786.30	19804.90	35267.97

NOTE: Standard errors are given in parenthesis. Also, the significant regression and dispersion parameters at $\alpha = 0.05$ are marked with an asterisk.

Table 22. Frequency comparisons for exercise data

Count	Observed	ZkICMP	ZICMP	ZkINB	ZkIP
0	3805	3805.21	3805.15	3804.32	3804.81
1	181	181.86	176.08	163.39	61.76
2	252	236.90	250.72	240.91	157.45
3	306	271.06	303.80	285.41	280.68
4	200	272.14	315.67	285.17	362.90
5	482	482.05	297.58	482.14	482.47
6	143	222.79	263.40	222.16	342.38
7	254	180.38	211.40	176.13	256.97
8	79	148.93	170.96	141.95	186.59
9	40	113.97	127.16	107.70	116.01
10	229	84.71	90.93	79.84	65.72
11-12	90	58.44	59.95	55.89	31.98
13-14	61	42.67	41.56	41.53	16.61
ABE		614.79	838.31	625.70	1219.90
χ^2		434.72	563.45	471.05	1054.30

CHAPTER 5

SUMMARY AND EXTENSIONS

5.1 SUMMARY

Count data occur frequently in a wide variety of scientific studies. The most popular model to analyze such data is the Poisson distribution. When the data consists of large number of observations which are zero, an appropriate model is the zero inflated Poisson (ZIP), which was made popular in a seminal paper by Lambert (1992). However, there are several situations where count data, besides zero, consists of high frequency for another positive count k. In this dissertation we examined two statistical models for such doubly inflated count data. First, we studied the zero and k inflated Poisson (ZkIP) model, which is an extension of ZIP. We discussed the distributional properties of ZkIP distribution, including a stochastic representation which facilitates parameter estimation. For grouped observations, we discussed two parameter estimation methods, maximum likelihood and expectation maximization (EM) algorithm. The elements of the Fisher information matrix to get the standard errors were also derived. We also studied in detail an alternative method, originally given by Louis (1982), to get the standard errors of the parameter estimates obtained using the EM algorithm.

For conducting test of hypothesis, we derived the asymptotic distribution of the likelihood ratio test statistic for testing the mixing probability on the boundary. The limiting distribution turns out to be a mixture of chi-square distributions with equal weights. For subject-specific count data that consists of covariates as well, we studied regression models that link the rate parameter of the Poisson distribution to the covariates. Other extensions where the mixing probabilities are allowed to depend on the covariates are straightforward. The methodologies that we studied have various applications in areas like manufacturing, transportation, econometrics, ecology etc. We have used two real life examples from health science to illustrate our

methods. The AIC, absolute error (ABE), and likelihood ratio (LRT) criteria were used to find the model that fits the given data.

The second part of this dissertation deals with zero and k inflated Conway-Maxwell-Poisson (ZkICMP) model. It is an extension of the ZkIP model. It is more flexible than ZkIP model as it not only captures inflation at zero and k but also the under and overdispersion that may be present in the count data. We studied the distributional and probabilistic properties of the ZkICMP distribution. The ZkICMP regression model is constructed to study the relationship between the explanatory variables and the count responses. We derived the maximum likelihood estimates and Fisher information matrix to get the standard errors of the unknown parameters. The ZkICMP could also be used for statistical analysis of count data with inflated frequencies. We have illustrated application of the ZkICMP on two count data examples from the National Health and Nutrition Examination Survey (NHANES).

In the next section we will discuss a brief overview of our ongoing research, possible extensions, and future research topics that are related to this dissertation.

5.2 EXTENSIONS

There are many possible extensions of the research in this dissertation that one could pursue. In this section we will describe our work in progress and future research problems that we intend to pursue.

5.2.1 ESTIMATION OF ZKICMP USING EM ALGORITHM

Maximum likelihood estimation of the parameters for ZkICMP could pose convergence problems, and the standard errors could be difficult to obtain. In Chapter 3 we described expectation maximization (EM) algorithm to get the ML estimates for ZkIP model. We also described on how to obtain the standard errors for the EM estimates using the method described by Louis (1982). We are currently pursuing extensions of these methods to the ZkICMP model. Here, we briefly outline the EM algorithm for the ZkICMP model. Let $\mathbf{u} = (u_1, u_2, u_3)$ be the vector of latent indicator variables for the three distributions in the mixture. Treating $\mathbf{u}$ as the missing

data, the likelihood function of the complete data $(\mathbf{y}, \mathbf{u})$ for the ZkICMP model is given by

$$\begin{aligned} L_{comp}(\pi_1, \pi_2, \boldsymbol{\lambda}, \nu|\mathbf{y}) &\propto \prod_{i:y_i=0} \left(\pi_1\pi_3 \frac{1}{Z(\lambda_i,\nu)}\right)^{u_{1i}} \prod_{i:y_i=k} \left(\pi_2\pi_3 \frac{{\lambda_i}^k}{(k!)^\nu Z(\lambda_i,\nu)}\right)^{u_{2i}} \\ &\quad \prod_{i:y_i\neq 0,k} \left(\pi_3 \frac{{\lambda_i}^{y_i}}{(y_i!)^\nu Z(\lambda_i,\nu)}\right)^{u_{3i}} \\ &\propto \prod_{i:y_i=0} (\pi_1\pi_3\, p_{0i})^{u_{1i}} \prod_{i:y_i=k} (\pi_2\pi_3\, p_{ki})^{u_{2i}} \prod_{i:y_i\neq 0,k} (\pi_3\, p_{y_i})^{u_{3i}} \end{aligned}$$

where $\pi_3 = (1 - \pi_1 - \pi_2)$, $p_{y_i} = {\lambda_i}^{y_i}/[(y_i!)^\nu Z(\lambda_i, \nu)]$ and $y_i \geq 0$. The loglikelihood function of the complete data is

$$\begin{aligned} \ell_{comp}(\boldsymbol{\theta}) = \log\, L_{comp}(\boldsymbol{\theta}|\mathbf{y}) &= \sum_{i:y_i=0} (u_{1i}\log(\pi_1) + u_{3i}(\log(\pi_3) + \log(p_{0i}))) \\ &\quad + \sum_{i:y_i=k} (u_{2i}\log(\pi_2) + u_{3i}(\log(\pi_3) + \log(p_{ki}))) \\ &\quad + \sum_{i:y_i\neq 0,k} (\log\pi_3 + \log p_{y_i}) \end{aligned} \tag{44}$$

where $\boldsymbol{\theta} = (\pi_1, \pi_2, \boldsymbol{\lambda}, \nu)$ is the unknown parameter vector. To implement the EM algorithm, we can replace the u_{ij}'s by their posterior means (E-step) and maximize equation (44) (M-step) to estimate $\boldsymbol{\theta}$. We can also obtain the standard errors using the method due to Louis (1982) outlined in Chapter 2. Implementation of the EM method for the ZkICMP regression model is straightforward.

5.2.2 ANALYSIS OF DOUBLY INFLATED COUNT DATA USING ANN

Recently, Haghani et al. (2017) analyzed the zero inflated count data using artificial neural networks (ANN). We plan to extend their ANN techniques to the ZkIP and ZkICMP models, for subject-specific as well as for grouped data. Also, we are interested in implementing ANN for both univariate and multivariate doubly inflated count data. The neural networks have been proved to be very reliable in many areas and implementing them for inflated count data will be an invaluable and cutting edge tool. The combination of classical methods and ANN will provide not only efficient results but also simple explanations leading to wider utility.

5.2.3 EXTENSIONS TO MULTIVARIATE ZKICMP

The focus of this dissertation has been on the univariate count responses. In a recent paper Sengupta et al. (2016) studied doubly inflated bivariate Poisson models for bivariate count response data. An extension of the bivariate Poisson model is the bivariate CMP given in Sellers et al. (2016). We are currently extending our results to the bivariate zero and k inflated CMP models. Our results can be regarded as a generalization of the paper by Sengupta et al. (2016). In general a multivariate CMP distribution can be constructed using copulas. We are in the process of extending our ZkICMP to the multivariate case using the Gaussian copula with ZkICMP marginals. These results will be generalization of the paper by Sen et al. (2017).

REFERENCES

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002), "Zero-inflated models with application to spatial count data," *Environmental and Ecological Statistics*, 9(4), 341–355.

Akaike, H. (1974), "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, 19(6), 716–723.

Alshkaki, R. S. A. (2016), "On the Zero-One Inflated Poisson Distribution," *International Journal of Statistical Distributions and Applications.*, 2(4), 42–48.

Atkins, D., and Gallop, R. (2007), "Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models.," *Journal of family psychology*, 21(4), 726–735.

Balakrishnan, N., and Pal, S. (2015), "An EM algorithm for the estimation of parameters of a flexible cure rate model with generalized gamma lifetime and model discrimination using likelihood- and information-based methods," *Computational Statistics*, 30(1), 151–189.

Barriga, G. D., and Louzada, F. (2014), "The zero-inflated Conway–Maxwell–Poisson distribution: Bayesian inference, regression modeling and influence diagnostic," *Statistical Methodology*, 21(Supplement C), 23–34.

Bohning, D., and Seidel, W. (2003), "Editorial: recent developments in mixture models," *Computational Statistics & Data Analysis*, 41(3), 349 – 357. Recent Developments in Mixture Model.

Cameron, A. C., and Trivedi, P. K. (2013), *Regression Analysis of Count Data*, London, UK: Cambridge Press.

Chant, D. (1974), "On Asymptotic Tests of Composite Hypotheses in Nonstandard Conditions.," *Biometrika*, 61(2), 291–298.

Choo-Wosoba, H., Levy, S. M., and Datta, S. (2016), "Marginal regression models for clustered count data based on zero-inflated Conway–Maxwell–Poisson distribution with applications," *Biometrics*, 72(2), 606–618.

Cohen, A. C. (1960), "Estimating the Parameters of a Modified Poisson Distribution," *Journal of the American Statistical Association*, 55, 139–143.

Conway, R. W., and Maxwell, W. L. (1962), "A queuing model with state dependent service rates," *Journal of Industrial Engineering*, 12, 132–136.

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society B*, 39(1), 1–38.

Dunn, P., and Smyth, G. (1996), "Randomized Quantile Residuals," *Journal of Computational and Graphical Statistics*, 5(3), 236–244.

Finkelman, M. D., Green, J. G., Gruber, M. J., and Zaslavsky, A. M. (2011), "A Zero- and K-Inflated Mixture Model for Health Questionnaire Data.," *Statistics in Medicine*, 30(9), 1028–1043.

Ghosh, S. K., Mukhopadhyay, P., and Lu, J.-C. (2006), "Bayesian analysis of zero-inflated regression models," *Journal of Statistical Planning and Inference*, 136(4), 1360 – 1375.

Greene, W. (1994), Accounting for Excess Zeros and Sample Selection in Poisson and Negative Binomial Regression Models, Working papers, New York University, Leonard N. Stern School of Business, Department of Economics.

Gupta, P. L., Gupta, R. C., and Tripathi, R. C. (1996), "Analysis of zero-adjusted count data," *Computational Statistics & Data Analysis*, 23(2), 207 – 218.

Gurmu, S., and Trivedi, P. (1996), "Excess Zeros in Count Models for Recreational Trips," *Journal of Business & Economic Statistics*, 14(4), 469–477.

Haghani, S., Sedehi, M., and S., K. (2017), "Artificial Neural Network to Modeling Zero-inflated Count Data: Application to Predicting Number of Return to Blood Donation," *J Res Health Sci.*, 17(3), E1–4.

Hall, D. B. (2000), "Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study.," *Biometrics*, 56(4), 1030–1039.

Kadane, J. B., Shmueli, G., Minka, T. P., Borle, S., and Boatwright, P. (2006), "Conjugate analysis of the Conway-Maxwell-Poisson distribution," *Bayesian Analysis*, 01(2), 363–374.

Lambert, D. (1992), "Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing," *Technometrics*, 34, 1–14.

Lin, T. H., and Tsai, M.-H. (2012), "Modeling health survey data with excessive zero and K responses," *Statistics in Medicine*, 32, 1572–1583.

Loeys, T., Moerkerke, B., De Smet, O., and Buysse, A. (2012), "The analysis of zero-inflated count data: Beyond zero-inflated Poisson regression.," *British Journal of Mathematical and Statistical Psychology*, 65(1), 163–180.

Lord, D., Guikema, S. D., and Geedipally, S. R. (2008), "Application of the Conway–Maxwell–Poisson generalized linear model for analyzing motor vehicle crashes," *Accident Analysis & Prevention*, 40(3), 1123 – 1134.

Lord, D., Washington, S. P., and Ivan, J. N. (2005), "Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory," *Accident Analysis & Prevention*, 37(1), 35 – 46.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society B*, 44, 226–233.

McLachlan, G. J., and Peel, D. (2000), *Finite mixture models*, New York: Wiley Series in Probability and Statistics.

Min, Y., and Agresti, A. (2005), "Random effect models for repeated measures of zero-inflated count data," *Statistical Modelling*, 5(1), 1–19.

Qin, X., Ivan, J. N., and Ravishanker, N. (2004), "Selecting exposure measures in crash rate prediction for two-lane highway segments," *Accident Analysis & Prevention*, 36(2), 183 – 191.

Ridout, M., and Besbeas, P. (2004), "An empirical model for underdispersed count data," *Statistical Modelling*, 4(1), 77–89.

Ridout, M., Demetrio, C., and Hinde, J. (1998), Models for count data with many zeros, in *International Biometric Conference, Cape Town.*

Rodrigues, J., de Castro, M., Cancho, V. G., and Balakrishnan, N. (2009), "COM–Poisson cure rate survival models and an application to a cutaneous melanoma data," *Journal of Statistical Planning and Inference*, 139(10), 3605 – 3611.

Saffari, S. E., and Adnan, R. (2011), "Zero-Inflated Poisson Regression Models with Right Censored Count Data," *Matematika*, 27(1), 21–29.

Salehi, M., and Roudbari, M. (2015), "Zero inflated Poisson and negative binomial regression models: application in education.," *Medical Journal of the Islamic Republic of Iran*, 29.

Sellers, K. F., Borle, S., and Shmueli, G. (2012), "The COM-Poisson model for count data: a survey of methods and applications," *Applied Stochastic Models in Business and Industry*, 28(2), 104–116.

Sellers, K. F., Morris, D. S., and Balakrishnan, N. (2016), "Bivariate Conway–Maxwell–Poisson distribution: Formulation, properties, and inference," *Journal of Multivariate Analysis*, 150(Supplement C), 152 – 168.

Sellers, K. F., and Raim, A. (2016), "A flexible zero-inflated model to address data dispersion," *Computational Statistics and Data Analysis*, 99, 68–80.

Sen, S., Sengupta, P., and Diawara, N. (2017), "Doubly-inflated Poisson model using Gaussian copula," *Communications in Statistics - Theory and Methods*, .

Sengupta, P., Chaganty, N. R., and Sabo, R. T. (2016), "Bivariate doubly inflated Poisson models with Applications," *Journal of Statistical Theory and Practice*, 10(1), 202–215.

Shapiro, A. (1985), "Asymptotic Distribution of Test Statistics in the Analysis of Moment Structures Under Inequality Constraints.," *Biometrika*, 72(1), 133–144.

Sheth-Chandra, M. (2011), The doubly inflated Poisson and related regression models, PhD Dissertation, Old Dominion University.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., and Boatwright, P. (2005), "A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution," *Royal Statistical Society: Series C (Applied Statistics)*, 54, 127–142.

Tang, Y., Liu, W., and Xu, A. (2017), "Statistical inference for zero-and-one-inflated poisson models," *Statistical Theory and Related Fields*, 1(2), 216–226.

Telang, R., Boatwright, P., and Mukhopadhyay, T. (2004), "A Mixture Model for Internet Search-Engine Visits," *Journal of Marketing Research*, 41(2), 206–214.

Umbach, D. (1981), "On inference for a mixture of a poisson and a degenerate distribution," *Communications in Statistics - Theory and Methods*, 10(3), 299–306.

Welsh, A., Cunningham, R., Donnelly, C., and Lindenmayer, D. (1996), "Modelling the abundance of rare species: statistical models for counts with extra zeros," *Ecological Modelling*, 88(1), 297 – 308.

Yang, Y., and Simpson, D. G. (2012), "Conditional Decomposition Diagnostics for Regression Analysis of Zero-inflated and Left-censored Data," *Statistical Methods in Medical Research*, 21(4), 393–408.

Yau, K., and Lee, A. (2001), "Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme," *Statistics in Medicine*, 20(19), 2907–20.

Zhang, C., Tian, G.-L., and Ng, K. (2016), "Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods," *Statistics and its interface*, 9(1), 11–32.

VITA

Monika Arora
Department of Mathematics and Statistics
Old Dominion University
Norfolk, VA 23529

Education

Ph.D	Old Dominion University, Norfolk, VA. (August 2018) Major: Computational & Applied Mathematics (Statistics).
MS	Old Dominion University, Norfolk, VA. (December 2015) Major: Computational & Applied Mathematics (Statistics).
M.Sc	Indian Institute of Technology-Bombay, Mumbai, India. (May 2011) Major: Applied Statistics and Informatics.
B.Sc	University of Rajasthan, Rajasthan, India. (May 2007) Major: Statistics, Mathematics, Computer Science.

Experience

Statistical Analyst, Chesapeake Bay Program, Old Dominion University, Norfolk, VA, (01/2014 - 07/2018).

Graduate Teaching Assistant, Old Dominion University, Norfolk, VA, (08/2013 - 12/2013).

Research Scientist, Covacsis Technologies Pvt. Ltd., Mumbai, India, (08/2011 - 07/2013).

Typeset using LaTeX.